MEASUREMENT
in Today's Schools

ROSS

LIBRARY OF PRINCETON
JAN 20 1954
THEOLOGICAL SEMINARY

LB1131
.R82

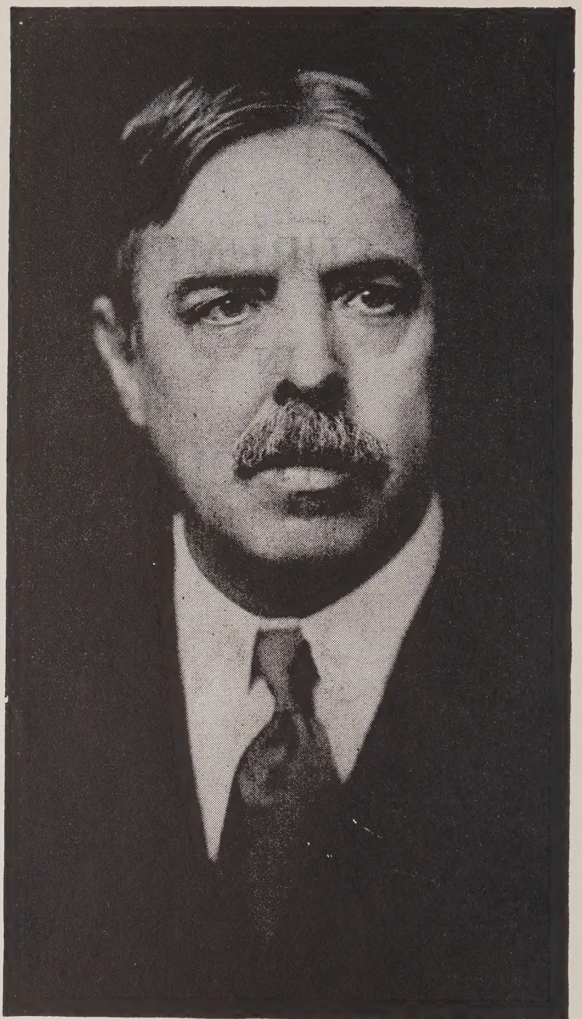
Digitized by the Internet Archive
in 2024 with funding from
Princeton Theological Seminary Library

https://archive.org/details/measurementintod00ross_0

MEASUREMENT

IN

TODAY'S SCHOOLS



Harris E. Ewing Photographic News Service

Edward Lee Thorndike

MEASUREMENT
IN
TODAY'S SCHOOLS

LIBRARY OF PRINCETON
JAN 20 1954
THEOLOGICAL SEMINARY

by

C. C. ROSS, Ph.D.

LATE PROFESSOR OF EDUCATIONAL PSYCHOLOGY
UNIVERSITY OF KENTUCKY

SECOND EDITION

New York
PRENTICE-HALL, INC.

PRENTICE-HALL PSYCHOLOGY SERIES

COPYRIGHT, 1941, 1947, BY

PRENTICE-HALL, INC.
70 FIFTH AVENUE, NEW YORK

ALL RIGHTS RESERVED. NO PART OF THIS BOOK MAY BE REPRODUCED IN ANY FORM, BY MIMEOGRAPH OR ANY OTHER MEANS, WITHOUT PERMISSION IN WRITING FROM THE PUBLISHERS.

First Printing	March, 1941
Second Printing	February, 1942
Third Printing	March, 1944
Fourth Printing	March, 1945
Fifth Printing	October, 1945
Sixth Printing	May, 1946
Seventh Printing	November, 1946

SECOND EDITION

First Printing	August, 1947
Second Printing	May, 1948
Third Printing	September, 1948
Fourth Printing	January, 1949
Fifth Printing	August, 1949
Sixth Printing	September, 1949
Seventh Printing	September, 1950
Eighth Printing	January, 1953

PRINTED IN THE UNITED STATES OF AMERICA

Preface to the First Edition

It is doubtless true that more progress in measurement has been made during the past quarter of a century than during all the years preceding. But the pattern of the measurement books has remained much the same. They have been very definitely centered about subject matter. The treatment has usually been organized around the conventional school subjects, and much space has been devoted to lists and descriptions of the measuring instruments available.

Authors of these texts have encountered obstacles increasingly difficult to surmount. The rapid increase in the number of tests and scales published has made it impossible to keep the books either complete or up to date. Even the most carefully compiled list of selected tests was likely to be rendered obsolete by the publication of better tests before the book was off the press. Fortunately, in recent years the appearance of rather complete and frequently revised bibliographies of published tests, together with critical evaluations, has made detailed lists and descriptions of available measuring instruments in textbooks no longer necessary.

Meanwhile, instructors in measurement have manifested a growing dissatisfaction with existing texts on the subject. For example, the typical class in measurement for high-school teachers has consisted of persons representing a variety of fields, but no one person has been interested in more than two or three of those discussed in the textbook, the rest of the material being largely deadwood. At the same time the enormous expansion of the experimental literature relating to measurement has had to be considered in any course that is at all adequate. And here the average book has left much to be desired.

Fifteen years' experience in teaching educational measurement to college classes has led the author to attempt a functional approach to the subject. The present work is the outgrowth of this experience. The emphasis is, therefore, not so much upon the description of the tools themselves as upon the multitude of problems relating to their intelligent use and interpretation by classroom teachers and school administrators.

It appears to the author that the time has come for a critical appraisal of measurement in today's schools, and for a careful search for generalizations to guide both theory and practice. The experimental evidence supporting these generalizations has been examined, and wherever possible reported in the language of the original author.

Since the functions of measurement are much the same on all educational levels, the illustrations have been drawn from both the elementary school and the secondary school, and to some extent from college. It is hoped that the book will be found useful to teachers, and to prospective teachers, regardless of the subject or the level of instruction.

In the preparation of the book the author has incurred obligations that are numerous and great. His first major indebtedness has been to his former teachers, notably Professors Edward L. Thorndike and William A. McCall, of Teachers College, Columbia University. The heavy obligation which the author owes to his co-workers in the field of measurement, upon whose publications he has freely drawn, is indicated by the numerous citations throughout the book. The fullest co-operation of these authors and their publishers is gratefully acknowledged. Special thanks are due to Professor A. B. Crawford, who has used a preliminary edition of the book at Transylvania College and at the University of Kentucky, and who has made numerous constructive suggestions; and to Professor G. M. Ruch, of the United States Office of Education, who has read the manuscript, and whose pertinent criticisms have been invaluable. Finally, the author is indebted to his own students, who for three years have used a preliminary edition of the book and who have offered many suggestions that have contributed greatly to its improvement.

C. C. Ross

Preface to the Second Edition

Since publication of the first edition of this book, experimentation in the field of measurement has made considerable progress. In this revision, the author has taken advantage of developments in this field, revising almost all the material from the first edition and including a great deal of altogether new material. All bibliographies and citations have been revised; a list of leading publishers of tests has been added to the final chapter.

Chapter tests and exercises, incorporated directly into each chapter of the first edition, have now been compiled into a separate workbook. This arrangement is designed to save valuable time for the student working on the exercises and for the teacher correcting them.

Sincere appreciation is due Mrs. Billy Whitlow Smith for considerable work both in the preparation of the manuscript and in correcting proof. Her assistance has been invaluable at every stage of the book's progress.

TO

PROFESSOR EDWARD LEE THORNDIKE

FOR NEARLY HALF A CENTURY
LEADER OF THE MOVEMENT
TOWARD A SCIENCE OF EDUCATION

Contents

PART I. THE PROBLEM OF MEASUREMENT

CHAPTER PAGE

I. Measurement in the Modern World 3

II. The Historical Development of Measurement in Education 27

III. The Characteristics of a Satisfactory Measuring Instrument 65

PART II. THE CONSTRUCTION OF INFORMAL TEACHER-MADE TESTS

IV. General Principles of Test Construction 103

V. Principles of Constructing Specific Types of Objective Tests 127

VI. The Construction and Use of Essay Examinations . 157

PART III. THE TESTING PROGRAM

VII. Steps in the Testing Program 175

VIII. The Statistical Analysis of Test Results 216

IX. The Graphical Representation of Educational Data 252

X. The Uses and Limitations of Norms 282

PART IV. MEASUREMENT IN INSTRUCTION

XI. Motivation 315

XII. Practice 346

XIII. Diagnosis 364

XIV. School Marks 397

XV. Classification and Promotion 423

XVI. Guidance 447

XVII. Evaluation 488

XVIII. Public Relations 515

INDEX 535

List of Illustrations

FIGURE PAGE

Edward Lee Thorndike . . . . *frontispiece*
Wilhelm Wundt 32
Karl Pearson and Sir Francis Galton 33
Alfred Binet 35
James McKeen Cattell 37
Lewis M. Terman 39
1. Test 7 from the Army Alpha 42
2. Test 6 from the Army Beta 43
3. A Scale for Measuring Pupils' Attitudes Toward High School 57
4. An Illustration of the Procedure Followed in Scoring Test 3 of the Terman Group Test of Mental Ability, Form A 203
5. A Sample Standard Test Scoring Record 204
6. An Educational Profile for a Standardized Achievement Test 210
7. Test Data Summary from the Cumulative Guidance Record of the Department of Supervision and Curriculum Development of the National Education Association . 212
8. A Cumulative Record in Graphical Form 213
9. An Effective Chart from a City Superintendent's Annual Report 253
10. Trends in Elementary-School Enrollments in Kentucky for a Six-Year Period 254
11. A Simple Bar Graph Which Shows Striking Differences Among the 48 States 255
12. A More Complex Bar Graph with High Attention Value 256
13. Profile of a Pupil and the Fifth-Grade Class of Which He Is a Member 260
14. The Profile of an Eleventh-Grade Pupil on the Progressive Achievement Test, Advanced Battery 261
15. A Profile for Representing Unit Scores in Achievement and Aptitude 262
16. A Profile of a College Student for Use in Guidance . . 263
17. A Histogram, or Column Diagram, Representing the Percentage Values Assigned to an Arithmetic Paper by Forty-Two Scorers 265

FIGURE PAGE
18. A Histogram, or Column Diagram, Representing the Distribution of IQ's in a Small Junior High School . . . 266
19. The Distribution of Mental Ages in an Eighth-Grade Class of Twenty-Seven Pupils 266
20. A Frequency Polygon Representing the Percentage Values Assigned to an Arithmetic Paper by Forty-Two Scorers 267
21. An Actual Curve Compared with the Theoretical Curve of Probability 267
22. A Percentile Curve Representing the Percentage Values Assigned to an Arithmetic Paper by Forty-Two Scorers 268
23. A Percentile Curve Representing the Distribution of IQ's in a Small Junior High School 269
24. Negative and Positive Skewness 270
25. Bar Graph Made on the Typewriter, Showing the Distribution of IQ's in a Small Junior High School . . . 270
26. Bar Graph Made on the Typewriter, Showing the Percentage of Pupils of Each Age Group Who Graduate from High School and the Percentage Who Enter High School but Do Not Graduate 271
27. Graph Made on the Typewriter, Showing the Overlapping of Grades Seven, Eight, and Nine in Reading Comprehension 272
28. Frequency Polygons Representing the Distribution in Reading Comprehension on the Iowa Silent Reading Tests for the Seventh, Eighth, and Ninth Grades of a Certain School 273
29. Total Comprehension Scores on the Iowa Silent Reading Tests for the Seventh, Eighth, and Ninth Grades . . 274
30. The Learning of Three Groups Compared, One with Full Knowledge of Progress, One with Partial Knowledge of Progress, and One with No Knowledge of Progress . 275
31. Correct and Incorrect Location of the Norms in a Line Chart, Showing Median Scores on a Reading Test . . 275
32. Grade Profiles for the Third, Fourth, Fifth, and Sixth Grades of a Certain School Made by Connecting the Median Scores on Each Part of the New Stanford Achievement Tests 277
33. A Line Graph Showing the Medians and Quartiles for Grades Four to Nine, Inclusive, in Reading Comprehension 278
34. The Central Tendency and Variability in Educational Age of Grades 2B to 9A, Inclusive, in a Small City School System 279

FIGURE PAGE
35. Percentages in a Typical Group Whose IQ's Fall Below and Above Various Points 299
36. The Relation Between Standard Scores, Percentile Scores, and Revised Stanford-Binet IQ's 300
37. The Profiles of Two Pupils Who Made the Same Total Score on a General Achievement Test 303
38. The Influence of a Knowledge of Progress upon Achievement in a College Class 333
39. A Study of the Influence of Praise and Reproof upon Achievement in Fourth-Grade and Sixth-Grade Arithmetic 338
40. Test-Machine and Drill-Machine Methods of Teaching Educational Psychology Compared with Ordinary Methods 357
41. The Five Levels of Educational Diagnosis 368
42. Analysis Sheet of Test 3, Metropolitan Achievement Tests, Form A, Arithmetic Fundamentals, for a Fifth-Grade Class in October 372
43. Traxler Chart of Suggested Diagnostic and Remedial Procedures in Handwriting 381
44. Areas in a Normal Distribution, Indicated by a Five-Letter Marking System Based on M and σ 415
45. General Quality of 200 Secondary Schools as Judged by Field Committees 425
46. Distribution of Mean Scores of Seniors in Forty-Nine Colleges in Pennsylvania on a Test of General Academic Knowledge 426
47. Distributions of Composite IQ's on Forms L and M of the New Revised Stanford-Binet Tests for a Standardized Group of 2,904 Individuals of CA's 2 to 18 Years . . 427
48. The Complete Scope of Guidance 451
49. Achievement Levels Corresponding to Various Levels of Adult Intelligence 458
50. Distribution of Scores on the Army Alpha for Five Occupational Groups 460
51. Shifts in Major Occupational Groups in the United States from 1870 to 1940 466
52. Distribution of Choices for Ten Occupations Made by 1,000 Boys in New York City Compared with the Number of Men Among 1,000 Workers Actually Following These Occupations in the City 467
53. A Suggested Technique for Evaluating the Philosophy of a Secondary School 499

FIGURE PAGE
54. Instructions for Using the Evaluative Criteria Developed by the Cooperative Study of Secondary School Standards 501
55. Summary of Evaluative Criteria for the Median Secondary School 502
56. An Evaluative Procedure for the Content of the Offerings in the Principal Subject-Matter Fields of a Secondary School 503
57. The Computation of Three Measures of the Adequacy of the Book Collections in the Library of a Secondary School 505
58. Evaluative Techniques for the Library Service of a Secondary School 506
59. The Computation of the Summary Score for the Guidance Service of a Secondary School 508
60. An Evaluative Technique for the Quality of Instruction in a Secondary School 509
61. A Suggested Informal Report to Parents 524
62. A Report Card Based on the University of Chicago High School 528

PART I

THE PROBLEM OF MEASUREMENT

CHAPTER I

Measurement in the Modern World

From birth to death almost every aspect of our daily lives is touched by measurement in its numerous forms. At birth the record of that important event is carefully made according to the nurse's watch. During the next few days measurements of the baby's weight and temperature are part of the daily routine of the hospital. Ever afterward, whether in school or outside, watches, clocks, balances, thermometers, money systems, and other forms of measurement play prominent roles in the life of every human being.

The daily round of the typical American probably begins somewhat like this: He rises at a certain hour by the clock, bathes in water measured by the meter, and dresses in clothing of a standard size. He begins his breakfast with half a grapefruit sold by the dozen, and sweetened with a spoonful of sugar sold by the pound. He continues with a bowl of cereal and two cups of coffee, both generously mixed with cream or milk sold by the quart. He then looks at his watch, jumps in his car, and watches the speedometer as he hurries to his work, for which he is paid by the hour, day, week, month, or year.

One day after another, year in and year out, he keeps this up until finally he worries himself to death over a falling stock market, measured in dollars, and a rising blood pressure, expressed in points. Then the hour of his departure is accurately noted, he is measured for a casket, and the time of his funeral is set according to the calendar and the clock. Afterward his life span is recorded in the family Bible and carved on his tombstone. In the meantime his estate is figured in dollars, and his widow lives the rest of her days from the income computed in per cent.

These common experiences are characteristic of the emphasis placed on measurement in the modern world. In fact, if all our various measuring devices were suddenly destroyed, contemporary civilization would collapse like a house of cards.

A. Measurement in Science

Apparently the chief problem of man has always been *adjustment*. As one writer puts it: "The civilization of a race is simply the sum-total of its achievements in adjusting itself to its environment."[1] The form of the problem has indeed varied somewhat from time to time, and still more has the method of meeting it. For ages the ingenuity of man was directed toward gaining practical *control* over the universe about him. At first the process was the uncritical procedure of trial and error. This fumbling way early led into such cul-de-sacs as alchemy, astrology, and magic. Later the seers and wise men began to attempt to put together these scattered bits of experience and so, in the words of Omar Khayyám, "To grasp this sorry Scheme of Things entire." Thus was born philosophy. The nature of the problem had then shifted to *understanding* the universe, rather than merely gaining control over it.

Scientific method. About three centuries ago there arose, with the experimental verification by Galileo of the laws of falling bodies, the method of modern science. Since that time man's quantitative conquest of nature has expanded not only into all branches of physics and chemistry but into organic and psychological phenomena as well. It is no exaggeration today to assert that science has revolutionized the material world in which we live. But it has done more than this; as Whitehead says, science has "practically recoloured our mentality."[2] As a distinguished chemist puts it: "Man's *inner* and *outer necessities,* real or imagined, have made him both a Scientist and a Philosopher."[3]

Both the *content* and the *method* of science are important. The content of science consists of a continuously expanding body of systematized knowledge, which is the product of scientific method. The one constant and universal feature of science is its method of arriving at knowledge.[4] John Dewey asserts that "the heart of science lies not in conclusions reached, but in the method of observation, experimentation, and mathematical reasoning by which conclusions are established."[5]

What, then, is the scientific method? Bertrand Russell suggests this concise formulation: "The essence of the scientific method is the discovery of general laws through the study of particular facts."[6] In another volume Russell elaborates this statement:[7]

---

[1] Hu Shih, "The Civilization of the East and the West," in *Whither Mankind,* edited by Charles A. Beard, page 27. New York: Longmans, Green & Company, 1928.

[2] Alfred North Whitehead, *Science and the Modern World,* page 3. New York: The Macmillan Company, 1925.

[3] Richard E. Lee, *The Backgrounds and Foundations of Modern Science,* page 3. Baltimore: The Williams & Wilkins Company, 1935.

[4] *Cf.* A. D. Ritchie, *Scientific Method,* page 14. New York: Harcourt, Brace & Company, Inc., 1923.

[5] *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II,* page 480. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1938.

> In arriving at a scientific law there are three stages: the first consists in observing the significant facts; the second in arriving at a hypothesis, which, if it is true, would account for these facts; the third in deducting from this hypothesis consequences which can be tested by observation.

But what is the role of measurement in scientific method? From Russell's analysis above, it would appear that, although measurement has but little, if any, bearing on the second stage in the scientific method, it is closely related to the first and third stages. Measurement performs a useful function in determining what alleged facts really are facts, as well as providing an exact method of describing them. It is also indispensable in the final stage of testing and verification, which is usually by means of specially devised experiments. A critical treatise[8] on educational measurement begins with this statement: "Measurement is the principal implement of science, changing that field of human endeavor from medieval gropings to a modern exactitude." The relationship is stated by Smart in the following words:[9]

> Of course, it must not be forgotten that our experience of sense qualities, in perception, serves as the basis of all scientific endeavor, and that this qualitative aspect of things is in varying degrees of completeness assimilated in and through the higher categories of the several natural sciences. And this assimilation is effected largely through the process of measurement, which thus functions as the connecting link between mathematics and the other sciences, and which is only a higher, i.e., more precise and complete form of that double-sided process of comparision and discrimination which begins on the qualitative level of experience.

Brief attention will now be given to the relation of measurement to each of the principal divisions of science. The discussion will observe the conventional divisions: namely, the "pure sciences" (physical, biological, and social) and the "applied sciences." Pure science is distinguished from applied science primarily on the basis of *purpose* or *motive,* and one division of pure science is distinguished from another on the basis of *subject matter.* Although the distinction is not always clear-cut, in general it may be said that pure science aims primarily at *understanding* the universe, whereas applied science aims at *predicting* and *controlling* it. The former will be considered first.

---

[6] Bertrand Russell, "Science" in *Whither Mankind, op. cit.,* page 65.

[7] Bertrand Russell, *The Scientific Outlook,* page 57. New York: W. W. Norton & Company, Inc., 1931.

[8] B. Othanel Smith, *Logical Aspects of Educational Measurement,* 182 pages. New York: Columbia University Press, 1938.

[9] Harold R. Smart, *The Logic of Science,* page 200. New York: D. Appleton and Company, 1931. Used by permission of D. Appleton-Century Company.

Measurement in the physical sciences. How can the place of measurement in any particular branch of science best be determined? It has seemed to the author that the chief reliance must be placed upon the testimony of outstanding scientists in the particular field and recognized historians of science. Astronomy is doubtless the oldest and among the most highly developed of the sciences. Although the rise of experimental science is usually dated from Galileo, who lived about 300 years ago, Boring describes two important experiments in astronomy which were made as far back as 2,200 years ago. In commenting upon these early experiments, Boring [10] says:

> It is no mere accident that these first two important astronomical experiments made use of mathematics in the interest of measurement. Measurement provides a precision of differentiation and definition in observation that can be had in no other way; mathematics provides the necessary means of carrying measurements through a logical development to their consequences without loss of their precision.

The bearing of this point upon the development of science is thus stated by Westaway: [11]

> The more that exact measurement enters into any branch of Science, the more highly is that branch developed. It is for this reason that Chemistry and Physics are so far in advance of Botany and Geology. And the reason why we can obtain so much clearer notions of, for instance, an area or a weight, than of, say, wisdom or chivalry, is because the former are *measurable,* the latter not. It is of the first importance in Science that we should, whenever possible, obtain precise quantitative statements of phenomena, and thus we see why it is that the introduction of a new scientific instrument so often leads to a marked advance in our knowledge.

Of the physical sciences, physics is usually regarded as the most highly developed at the present time. Two outstanding figures in the development of modern physics are Lord Kelvin in England and Max Planck in Germany. Regarding the place of mathematics and measurement in physics, Lord Kelvin says: [12]

> When you can measure what you are speaking about and express it in numbers, you know something about it, and when you cannot measure it,

---

[10] Edwin G. Boring, *A History of Experimental Psychology,* pages 14–15. New York: The Century Company, 1929. Used by permission of D. Appleton-Century Company.

[11] F. W. Westaway, *Scientific Method: Its Philosophical Basis and its Modes of Application,* pages 271–272. New York: Hillman-Curl, Inc., 1937.

[12] Quoted by Ronald King, "Physics, Metaphysics and Common Sense," *Scientific Monthly,* 42: 311, April, 1936

when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind. It may be the beginning of knowledge, but you have scarcely in your thought advanced to the stage of science.

Max Planck, regarded as one of the leading exponents of the important quantum theory in physics, declares that further progress in the physical sciences "will depend essentially on the development and wider application of our methods of measurement."[13] The case for measurement in physics may very well rest upon the testimony of these expert witnesses, although practically every prominent physicist from Galileo and Kepler to Einstein could be called.

Measurement in the biological sciences. The history of the various branches of science indicates clearly that not only are the biological sciences younger than the physical sciences, but also that measurement and mathematics have always occupied and still occupy a less prominent place in the biological sciences. The primary reason for this is doubtless the greater simplicity of the data in the physical sciences.[14]

Certainly, however, biology has progressed far beyond the stage of authority that existed in the Middle Ages, when, for example, the question of the number of teeth possessed by the horse was the subject of heated debate in many contentious writings. "Apparently," says Locy, "none of the contestants thought of the simple expedient of counting them, but tried only to sustain their position by reference to authority."[15] Locy recognizes three somewhat overlapping phases, or stages, in the development of the biological sciences: namely, the descriptive, the comparative, and the experimental.[16]

Without doubt, one of the most important generalizations of modern science is that of evolution through natural selection as set forth by Charles Darwin near the middle of the last century, and yet it will be remembered that his work was nonmathematical, consisting largely of the classification of vast amounts of data upon which these epoch-making generalizations were based. In fact, as far as method goes, it was largely an extension and refinement of that emphasized by Aristotle more than two thousand years earlier.

---

[13] Max Planck, *Where Is Science Going?*, page 96. New York: W. W. Norton & Company, Inc., 1932.

[14] Julian Huxley makes this statement: "Sciences, like Empires, have their rise and their time of flourishing, though not their decay. Naturally, the order of their rise runs parallel with the complexity of their subject-matter. The physical sciences, being the simplest and most straightforward, were the first to start their triumphant career." *What Dare I Think?* page 1. New York: Harper & Brothers, 1931.

[15] William A. Locy, *Biology and Its Makers* (Third Edition, Revised), page 18. New York: Henry Holt & Company, 1922.

[16] *Ibid.*, page 443.

Whitehead attributes the retardation of science in the Middle Ages largely to Aristotle's emphasis on classification rather than measurement. Note this statement: [17]

> But the biological sciences, then and till our own time, have been overwhelmingly classificatory. If . . . only the schoolmen had measured instead of classifying, how much they might have learnt!

That Darwin's cousin, Sir Francis Galton, took this position is clearly indicated by the following statement: "Until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science." [18] Galton, therefore, proceeded to introduce exact measurement and mathematical calculation into the theory of evolution. In later years these biometrical methods have been greatly extended by Karl Pearson, Spearman, and others. The remarkable experiments of Mendel in heredity appeared at about the same time, although their value was not recognized until about 1900. Owing to these pioneers, the knowledge of heredity has become established on a definite mathematical basis.[19] Even earlier than Mendel's time such noted physiologists as Müller, Weber, and Helmholtz had been doing much the same for physiology.[20] After a survey of the development of natural science from Aristotle to Fabre, which shows a definite trend from qualitative to quantitative analysis, Peattie concludes: "In short, what science calls for today are life histories, and ecological studies—the precise measurement of the environmental factors and the inter-relations of organisms." [21]

The various biological sciences have not been placed upon as definitely a quantitative basis as have physics and chemistry, largely because of the nature of their data. Some competent students of science think the biological sciences have moved too far in that direction. Whitehead,[22] for example, expresses regret that "biology apes the manners of physics," while at the same time neglecting the unique character of its own subject matter, organisms, which are incapable of analysis without the destruction of their essential nature. The Gestalt school of psychology in recent years has also registered a vigorous protest against the atomic conceptions of mind which experimental psychology took over from nineteenth-century physics.

---

[17] Alfred North Whitehead, *op. cit.*, page 43.

[18] Sir Francis Galton, quoted by I. W. Howerth in "Measurement of Mental Phenomena," *Phi Delta Kappan*, 15: 1–9, June 2, 1932.

[19] Erik Nordenskiöld, *The History of Biology*, pages 583–594. New York: Alfred A. Knopf, Inc., 1928.

[20] William A. Locy, *op. cit.*, page 192.

[21] Donald Culross Peattie, *Green Laurels*, page 345. New York: Simon and Schuster, Inc., 1936.

[22] Alfred North Whitehead, *op. cit.*, page 150.

Measurement in the social sciences. Measurement in the social sciences presents a difficult problem. The social sciences are not only newer than the natural sciences but their data are more complex. They study human beings, the most complex of all biological organisms, and their social relationships, which are more complex than purely individual responses.

The genetic history of the social sciences has been described[23] as follows:

> In the days of Aristotle, Plato, and Pythagoras, philosophy still embraced the exact, natural, and social sciences. At the beginning of the nineteenth century the exact and natural sciences—mathematics, astronomy, physics, chemistry, geology, biology—had already left their philosophical matrix and were rapidly developing their own methods and techniques, while preserving a tendency to return to philosophy for an occasional theoretical and speculative rehauling. But the social sciences—history, ethics, law, economics, psychology, religion, esthetics, anthropology (such as it was)—were still rocking in the metaphysical cradle of Mother Philosophy. One by one the babes emerged and learned to stand on their own feet and to talk their own language, even though their gait and vocabulary continued for a long time to bear traces of their maternal heritage.

In the preface to *The Seven Seals of Science,*[24] Mayer states his major thesis as follows:

> The central theme of the essay is that the sciences did not arise and could not have arisen simultaneously, that they form a well defined structure with mathematics at the bottom, that each later science built upon those that went before, that psychology is only now in process of becoming established, and that the social studies, if they are to be worthy of the name of science, must build upon the natural sciences and particularly upon geology, biology, and psychology.

It is significant that the author, although a professor of economics and sociology, uses as title for the final chapter in the book, "Social Science in the Making."

Other writers take a somewhat more optimistic position, and many of them indicate specifically the direction the development of social science is taking and must take. For example, Ogburn and Goldenweiser[25] write as follows:

> Attention, finally, must be drawn to the increasing importance of statistical methods in the social sciences. . . . The extent to which social thought and

---

[23] William F. Ogburn and Alexander Goldenweiser, *The Social Sciences and Their Interrelations,* pages 2–3. Boston: Houghton Mifflin Company, 1927.

[24] Joseph R. Mayer, *The Seven Seals of Science,* page vii. New York: The Century Company, 1927. Used by permission of D. Appleton-Century Company.

[25] William F. Ogburn and Alexander Goldenweiser, *op. cit.,* pages 8–9.

> theory will pass from the sphere of opinion, conjecture, and contemplative analysis to that of fact, knowledge, and control, will depend on their permeation by these scientific methods of measurement and statistics.

Of course, there is nothing particularly new about this viewpoint. As early as 1798 Malthus attempted to put economics on a definite mathematical basis when he announced his celebrated, although erroneous, proposition that "population increases in a geometrical ratio, food in an arithmetical ratio." A little later, in 1835, Quetelet showed that the theory of probability could be applied to human problems such as insurance.

Barnes traces the history of sociology and concludes: "There is a general agreement that sociology can become a true science of society only in the degree to which it is able to appropriate and apply those exact methods of measurement and analysis which constitute the indispensable attributes of science in general." [26] On the other hand, Ellwood,[27] an eminent sociologist, takes a wholly different position. His point of view is clearly stated [28] in these words:

> It would seem to me that as we ascend in the scale of life the view that science is quantitative measurement of objective conditions becomes less and less applicable, not only because measurement becomes more difficult, but because the subjective element plays a larger part. Even if the subjective element is capable of certain measurements and even if it is true that whatever exists exists in some quantity or number, nevertheless, it is obvious that where subjective elements play a large part, measurement becomes of less importance for accurate knowledge because it is confined to the superficial aspects of the total situation and fails to expose the nature of the process which is being investigated. This is especially true in the social sciences and in them measurement seems to me to play a rôle secondary to other scientific methods.

It seems fairly clear, therefore, that measurement and statistical analysis of quantitative data do occupy a prominent place in the social studies, although there is no general agreement as to just what this place is. There does appear, however, to be universal recognition that the problems are more difficult than those presented by the earlier sciences, and that their solution must be based at least in part on these other sciences, notably psychology. Measurement in psychology will be considered at some length in later sections.

Measurement in the applied sciences. It has already been pointed out that the distinction between pure and applied science is not always easy to draw. In the beginning, science appears to have arisen in the service of certain basic human needs and desires.[29] Despite its humble origin, however, science soon ceased to be but a means to an end and became an end in itself. For about a century and a half following Galileo it became exclusively the pursuit of the learned, and affected not at all the thoughts or habits of ordinary men. The emphasis had shifted from applied to pure science. Russell comments [30] upon this fact as follows:

---

[26] Harry Elmer Barnes, "The Development of Sociology," *Scientific Monthly,* 35: 547, December, 1932.

[27] Charles A. Ellwood, "The Uses and Limitations of the Statistical Method in the Social Sciences," *Scientific Monthly,* 37: 353–357, October, 1933.

[28] *Ibid.,* page 353.

> It is only during the last hundred and fifty years that science has become an important factor in determining the everyday life of everyday people. In that short time it has caused greater changes than had occurred since the days of the ancient Egyptians. One hundred and fifty years of science has proved more explosive than five thousand years of prescientific culture.

The cycle is now complete. Science, which arose from man's stern necessity for meeting the ordinary problems of life, has now returned to serve again his practical needs. And nowhere is measurement more in evidence than in its practical applications. A competent observer [31] asserts that "one can hardly think of a field of intellectual endeavor into which measurement has not crept, and surely there is none in which its influence has not been felt."

Indeed, it is these practical applications of science that have impressed the mind of the layman in recent years. When he hears the word "science," he is likely to think of the results of science, possibly some invention such as the radio or radar, rather than of the physics and chemistry that have made them possible. Probably a thousand persons would know of Marconi, who invented wireless telegraphy, to one who ever heard of Hertz or Maxwell, whose pioneer work blazed the trail.

The prominent place of quantitative measurement in these modern applications of science to engineering and industry is too well known to require elaboration. Take a modern automobile as an example. Its mechanical parts are accurate to the thousandth part of an inch. Every detail has been subjected both to careful laboratory experimentation and to rigid tests on the trial grounds. The instrument board presents, as practical aids to the user, various devices for measuring gas, electric current, oil pressure, temperature, and speed of car, as well as perhaps a clock and a radio.

It is instructive to study what the application of science has done for modern cookery. The ordinary untrained housewife still uses recipes with such vague directions as "season to taste," "add butter to the size of a walnut," "cook in a moderate oven," and so on. In contrast, the modern bakery accurately measures all ingredients, mixes them uniformly for a specified length of time, and then cooks them at a specified temperature for a definite time. This assures a predictable uniformity in the product, in contrast with the "luck" of the old-time cook.

---

[29] *Cf.* John Dewey, *How We Think,* page 216. Boston: D. C. Health & Company, 1933.

[30] Bertrand Russell, *The Scientific Outlook,* pages vii–viii.

[31] B. Othanel Smith, *op. cit.,* page 2.

Medicine is an outstanding field in which many discoveries of pure science have been applied to the solution of practical human problems. Dr. Herrick [32] of Chicago explains how the development and use of various instruments of precision have revolutionized medical diagnosis and practice. The measurement of blood pressure, body metabolism, and the physical and chemical analyses of the blood and other body fluids are as recognized techniques today as were height, weight, and temperature a generation ago. The dietitian in the kitchen measures the patient's food from the standpoint of calories, minerals, and vitamin content with an accuracy approaching that of the pharmacist in compounding his medicines and of the nurse in administering them.

Limitations of measurement. Before concluding this discussion of the relation between measurement and science, it may be well to note some of the difficulties and problems of measurement. It must not be assumed that the tools and techniques of measurement have been developed to a state of perfection. This is far from true even in physics and chemistry, where measurement has progressed furthest.

Planck, one of the most eminent physicists of the day, offers the warning that "every number obtained by physical measurements is liable to a certain possible error." [33] Westaway puts the matter in these words: "We may, in fact, look upon the existence of error in all measurements as the normal state of things." [34] Doubtless, Bertrand Russell has the same idea in mind when he describes science as a "succession of approximations." [35]

In general, it may be said that the sources of error in measurement are due to the imperfections either in the measuring instruments themselves, or in the method with which they are employed. While both of these sources of error are subject to a considerable measure of control, neither can be eliminated altogether. Three methods of controlling errors in measurement may be suggested:

---

[32] James B. Herrick, "Changes in Internal Medicine Since 1900," *Journal of American Medical Association,* 105: 1312–1315, October 26, 1935.

[33] Max Planck, *A Survey of Physics,* page 92. New York: E. P. Dutton & Co., Inc., 1925.

[34] F. W. Westaway, *op. cit.,* pages 289–290.

[35] Bertrand Russell, *The Scientific Outlook,* page 65.

1. The improvement of existing measuring instruments.
2. The devising of adequate methods of estimating or allowing for errors.
3. The development of skill in applying the instruments of measurement so as to reduce errors to a minimum and in interpreting the results so as to take due account of the errors which cannot be eliminated.

The first of these methods will be considered at some length in Chapters III to VI, the second will receive attention in Chapter VIII, while practically the entire book is concerned with the third.

It is a major thesis of the author that the limitations of existing measuring instruments do not detract one whit from the importance of measurement, although they do add to its difficulty. The result is rather to set a special premium upon the skillful use of these instruments. As a rule, the cruder any tool is, the greater the skill required in its application, if satisfactory results are to be obtained. The early automobile, for example, called for much greater skill in its successful operation than does its more highly perfected modern successor.

Conclusions. What, then, is the relation between measurement and science? A few generalizations seem fairly clear:

1. There is a direct relationship between the status of a science and the degree to which measurement has been developed in it. In the older and better established physical sciences, measurement occupies a fundamental place; in the newer biological sciences, measurement occupies a less important place; and in the social sciences, the most recent group, measurement has made hardly more than a beginning. The evidence seems abundantly to support Westaway's statement:[36] "The more that exact measurement enters into any branch of Science, the more highly is that branch developed."
2. The prominence of measurement in a science appears to be roughly in inverse ratio to the complexity of its subject matter. Inert material seems inherently more susceptible to measurement than living organisms. Apparently the maximum difficulty comes in the case of man, particularly in his social behavior.
3. All measurement is subject to errors. These errors are due to limitations in the tools as well as in the techniques of measure-

---

[36] F. W. Westaway, *op. cit.*, pages 271–272.

ment. To ensure satisfactory results, the greater the limitations of the former the greater the skill and insight called for in the latter.

B. Measurement in Education

Is education a science? Education is described sometimes as a philosophy, sometimes as a science, and sometimes as an art. Moreover, it is commonly assumed that these terms can be clearly distinguished from each other or even that there is a certain antagonism among them. Quite the contrary is the truth, however. They are most closely interrelated. As the role of measurement in education will doubtless be different if education is a science from what it will be if education is a philosophy or an art, some attention must now be given to the problem of the relationship of science to philosophy on the one hand and to art on the other.

Perhaps the kinship of science and philosophy will be more apparent if approached genetically and historically. The early Greek thinkers, Plato and Aristotle for example, recognized no distinction between science and philosophy, both being joined by the common bond of love, the pure love of truth.[37] During the Middle Ages, however, this once harmonious family found itself unequally and somewhat unhappily yoked together. Since the later Renaissance, one after another of the children born to this union, beginning with physics and chemistry, left the family tree and set up in business for themselves. This was a sort of "psychological weaning," which doubtless performed a useful function for the time being. But, as often happens when discontented youth desires to assert its independence, science rebelled altogether and would have nothing at all to do with the parental wisdom of Mother Philosophy. This needlessly prolonged period of arrogant adolescence has continued to the present time. Fortunately, in recent years some of the wiser heads have somewhat patched up the family quarrel which has brought unhappiness to mother and daughters alike.[38] A recent writer describes the result in education:[39]

> Educators are so over-zealous to become "scientific" and objective that they have ignored the fact that there is no less need for thinking in science than in philosophy. When research procedure neglects theoretical evaluation and interpretation, it is only partly scientific.

---

[37] Harold R. Smart, *op. cit.*, Chapter 1.

[38] For an excellent general discussion of this point of view, see: Max Black, "A Lend-Lease Program for Philosophy and Science," *Scientific Monthly,* 61: 165–172, September, 1945. For an admirable statement of the dangers of a narrow conception of a science of education, see: William A. Brownell, "Quantitative Research on Teaching and Learning," *School and Society,* 50: 847–856, December 30, 1939.

[39] J. Stanley Gray, "A Neglected Phase of Educational Research," *Journal of Educational Research,* 29: 89–90, October, 1935.

Although competent students of education recognize that both philosophy and science are necessary for a complete act of thinking, each field is so broad as to make a certain division of labor necessary. The situation has been well stated [40] as follows:

> If one specializes in the critical examination of educational theories, hypotheses, and generalizations in the light of data which are already available, we call him an educational philosopher. If one specializes in the solving of educational problems by making new appeals to experience through systematic, controlled and uncontrolled observation, in field or laboratory, we call him an educational scientist in the classical sense of the term.

But even the most competent philosopher is forced to take his science at second hand from the data made available by specialists in science, for it is too much to expect him at the same time to be an expert experimentalist.[41] Conversely, a competent scientist is forced to take his philosophy largely at second hand while he conducts his persistent search for new facts. Moreover, not only does a scientist have to borrow his philosophy, but much of his science also.

This borrowing, though necessary, is risky business not only for the philosopher but for the scientist as well.[42] Not only may he be unfortunate in what he borrows, but its meaning to him is inevitably colored by the mental background imposed by his specialty. The important point is that science and philosophy are reciprocally related, as inseparably linked as are heredity and environment in the growth of a living organism. Buckingham, a recognized leader in educational research, states the relationship concisely: "As fields of human endeavor science and philosophy supplement each other. . . . Without philosophy, science is incomplete; without science, philosophy is barren." [43] An educational philosopher puts the relationship as follows: "While philosophy must be the general to plan the grand strategy of education, it will need science as its staff officer." [44] Broadly conceived, education then is both science and philosophy. As a science it belongs to the group known as the social sciences, whose data are the most complex of all. While education is not, and doubtless never will be, as thoroughgoing a science as physics or chemistry, it has nevertheless made more progress toward the scientific treatment of its problems during the twentieth century than in all the centuries preceding.

---

[40] Carter V. Good, A. S. Barr, and Douglas E. Scates, *The Methodology of Educational Research,* page 24. New York: D. Appleton-Century Company, 1936.

[41] In commenting upon William James, Boring makes this disturbing observation: "It is too bad, but no one has ever yet succeeded in being both a good philosopher and a good experimentalist." E. G. Boring, *op. cit.,* page 502.

[42] *Cf.* Alfred North Whitehead, *Nature and Life,* page 5. Chicago: University of Chicago Press, 1934.

[43] B. R. Buckingham, "The Philosophy and Organization of Research," *School and Society,* 29: 758, June 15, 1929.

[44] John S. Brubacher, *Modern Philosophies of Education,* page 87. New York: McGraw-Hill Book Company, Inc., 1939.

But what of art in relation to science and philosophy? Will Durant puts the matter clearly and graphically: [45]

> Every science begins as philosophy and ends as art; it arises in hypotheses and flows into achievement. Philosophy is . . . the front trench in the siege of truth. Science is the captured territory; and behind it are those secure regions in which knowledge and art build our imperfect and marvelous world.

William H. Payne more than half a century ago suggested that "in the slow but sure evolution of human opinion, a science of education is beginning to emerge from the art of education." [46] But there is also a reciprocal relationship, for as Doughton [47] points out, "the sure foundation of an effective teaching art is a science of education."

Science consists in *knowing,* while art refers to *doing* and implies skill and aesthetic excellence. An outstanding teacher might be either an artist or a scientist, although the ideal teacher must be something of both. No matter how great the artist or with how much inspiration he wields the brush, the pigments are mixed according to formula.

The place of measurement in education. What, then, is the rightful place of measurement in education, which is at once a science, a philosophy, and an art? The answer varies somewhat with the point of view of the observer. Naturally the role of measurement appears more important to those educators whose specialty is science than to those whose specialty is philosophy. At times these views become so divergent as to appear wholly irreconcilable. The quotations on the opposite page, from outstanding educational leaders in a single institution, will make this clear.

It is always dangerous, of course, to detach a statement from its context. However, the strong language employed, containing such terms as "thoroughly," "indispensable," "final," "fallacy," "never," and "always," scarcely leaves room to hope that these statements are entirely harmonious. Doubtless, any attempt to interpret the above quotations should take into consideration the date of publi-

---

[45] Will Durant, *The Story of Philosophy,* pages 2–3. New York: Simon and Schuster, Inc., 1926.

[46] William H. Payne, *Contributions to the Science of Education,* page 11. New York: Harper & Brothers, 1886.

[47] Isaac Doughton, *Modern Public Education, Its Philosophy and Background,* page 308. New York: D. Appleton-Century Company, 1935.

TWO VIEWS OF MEASUREMENT IN EDUCATION

Educational Scientist	Educational Philosopher
Whatever exists at all exists in some amount. To know it thoroughtly involves knowing its quantity as well as its quality.[48]	Yet another false concept in the climate gripping the American scholar thwarts his study of Man. This is the fallacy that "I know only what I can describe quantitatively . . . whatever exists, exists in some measurable amount."[49]
Measurement is indispensable to the growth of scientific education. . . . The final answer to every educational question, except one, must be left to the educational measurer and must await the development of education as a science.[50]	And I should myself like further to conclude that education can never become a science . . . always—so long as this world stands—will there be problems, nay legions of problems, with which the processes of "exact" science are insufficient to cope.[51]

cation. It will be noted that views of the educational scientists represent an earlier period. They appeared soon after World War I, when the success of the Army Alpha test was still fresh in the minds of psychologists. Moreover, at that time, atomic physics still dominated all science, including psychology, and behaviorism was in the ascendancy in America. The more recent quotations from the educational philosophers reflect a newer atmosphere. The recognition of the principle of indeterminancy in the new physics has had a sobering effect on science in general, and the Gestalt school of psychology in particular has made a vigorous protest against what they regard as the atomic conception of mind. It is, therefore, quite possible that the present views of its two groups are much closer together than the above statements would indicate. However, the following statement[52] from McCall strongly suggests that not all differences have been ironed out:

> Certain extreme exponents of the organismic (often called *Gestalt*) view contend that any organism is more than the sum of its parts, and that adding test scores is like trying to make a man by sticking together a head, a trunk, two arms and two legs. But a reading score cannot be properly compared

---

[48] Edward L. Thorndike, *Seventeenth Yearbook of the National Society for the Study of Education, Part II,* page 16, 1918.

[49] Harold Rugg, *The Social Frontier,* 1: 12, March, 1935.

[50] William A. McCall, *How to Measure in Education,* pages 7, 9. New York: The Macmillan Company, 1922.

[51] William H. Kilpatrick, *School and Society,* 30: 48, July 13, 1929.

[52] *The Test Newsletter,* published by the Bureau of Publications, Teachers College, Columbia University, December, 1936.

> to one leg. It is not a broken off fragment of the mind. In a very real sense, a reading score tends to measure the entire organism functioning in that reading situation.
>
> Mental measurements are essentially similar to bodily measurements. If anyone proposed to abolish the making and use of the measurements of pulse, temperature, blood pressure, et cetera, we would call him *crazy,* and if anyone proposes to abolish the making and use of mental measurements, he, too, should be called—I hesitate to say what, since somehow I must manage to live with certain of my colleagues after this is published, but surely something other than an organismic philosopher or a *Gestalt* psychologist!

Both scientists and philosophers attempt to test their generalizations before finally accepting them. With science the process is the straightforward one of subjecting all such generalizations to rigid mathematical or experimental verification. With philosophy the process appears more involved. Kilpatrick,[53] for example, recognizes two distinct situations: "simple prophecies" and "decisions on appropriate conduct or policy." For the former, he agrees that "'verification' is an appropriate term and measurement (when available) is a proper means of testing." For the latter, however, he insists that "'verification' is not an appropriate term and techniques of measurement are not in themselves adequate." He continues: "In such cases the function of measurement is not to supplant or to supply decisions, but to furnish, regarding the working of the policy under review, more and better data, in the light of which a fresh and better decision can be made." Apparently, then, whenever actual verification is possible, this philosopher at any rate is willing to assign to measurement the job of doing it; and even in the other cases he assigns to it the necessary, if humble, duty of providing at least part of the data required. Perhaps, then, it is not an unfair statement to say that the scientist *always* assigns to measurement a fundamental role, whereas the philosopher *sometimes* does so. However, at all times the philosopher seems willing to ascribe to measurement an important, even if not a fundamental, place in education.

Take the important matter of guidance, for example. Even its most enthusiastic supporter would hardly characterize guidance as a full-fledged science. Yet measurement provides some of the essential data in any sound guidance program. To describe a pupil as of weak scholarship and of low mentality is to leave his status vague and unsatisfactory. But to say that he has a percentile rank of 20 on the Progressive Achievement Tests and an IQ of 84 on the Revised Stanford-Binet Tests of Intelligence is to describe him in precise and meaningful terms.

---

[53] William H. Kilpatrick, "The Relation of Philosophy to Scientific Research." *Journal of Educational Research,* 24: 110–114, September, 1931.

In this connection it is well to observe that measurement is always a means to an end, and never an end in itself. A measurement is simply a quantitative description of observed data. The significance or educational implications of the measurement are rarely self-evident or automatic. As a rule, the true significance of the measurement can only be determined when it is seen in relation to other relevant factors and is fitted into the total pattern of the situation. The term *evaluation,* as distinguished from *measurement,* is often used to refer to the process of appraising the whole child or the entire educational situation.

The three R's in education. Everyone is familiar with the famous trinity of R's in education, "Readin', 'Ritin', and 'Rithmetic." These, of course, have to do with the content of education, the curriculum of instruction. There is also another series of R's which is concerned with the process of arriving at, or at least of searching for, truth in education to serve as a basis for theory and practice. Educators have, in general, employed three principal methods of settling educational issues and of arriving at educational principles and policies. These constitute a new series of R's, "Rhetoric, Reputation, and Research."

Historically the first of these methods may be termed that of Rhetoric. It is the method *par excellence* of politicians, although unfortunately not unknown in education, especially among the reformers of every period. The method is usually most dangerous when used orally. It is too well known to require detailed discussion here. Abe Martin's famous definition of an orator as a "public speaker not unduly hampered by the facts" indicates rhetoric's limitations. The danger is that the personality of the speaker may outweigh the merits of the case, and the artistic form of the speech may have more influence than its content.[54] Naturally all measurement and quantitative data are irrelevant, if not positively in the way. As a matter of fact, a speaker of the House of Representatives attributed the decline in oratory chiefly to the "general diffusion in knowledge," since "as a rule the more information a man has the less emotional he is, and the orator's appeal was to the emotions far more than to the understanding."[55]

The second method of determining educational theory and practice may be termed that of Reputation. According to it, the settlement of an educational issue is the simple matter of finding out what has been said on the question by some persons whose reputations in the field are sufficiently great to make them accepted as authorities. This method has been the dominant one in education until recently and is still widely used. It has a legitimate and necessary place in education as it does in law and medicine, where one must rely for the solution of many practical problems upon the professional judgment of acceptable authorities. But the method is not without its dangers, which are so important as to warrant brief discussion.

---

[54] Irvin S. Cobb described a prominent Southern orator as one who can "make a song of a syllable and turn any reasonably long word into an anthem." *The Courier Journal,* Louisville, Kentucky, June 16, 1936.

[55] Champ Clark, "Is Congressional Oratory a Lost Art?" *Century,* 81: 310, December, 1910.

In the first place, the authority may be mistaken. Reputation is unfortunately no guarantee of reliability. Until comparatively modern times the wisest persons were quite certain that the sun each day made a complete journey around our flat earth. Such divergent views on practically every phase of education as are expressed in current educational journals and in public addresses by our most eminent educational leaders are ample assurance that men of the highest reputation may be mistaken.[56] In the second place, the authority may be misquoted. A few years ago at a meeting of the American Psychological Association a speaker quoted what purported to be a statement from an outstanding psychologist. At the conclusion of the address, this psychologist arose to explain that he had never made any such statement and in fact believed quite the contrary. It is too much to expect, however, that the authority will always be present to correct his alleged quotations. A third danger is that conditions may have changed so greatly that a statement once true may be no longer applicable.[57] For example, George Washington's warning against foreign entanglements, made in 1797, when the United States consisted of 16 states whose total population, barely 5,000,000, was separated from Europe by a broad Atlantic, need not be at all applicable to a nation of 48 states, with a population of more than 125,000,000, joined to Europe by modern agencies of communication. The method is thus seen to be beset by many dangers. It must, therefore, be used with caution. The necessity of extreme care in the selection of the authority cannot be overemphasized. It is usually wise, also, to examine the evidence that lies behind the statement and the conditions under which it was made. Reputation alone must not be thought of as adequate assurance of reliability. At best the reliance upon reputation involves considerable risk. It is always well to consider the circumstances under which the statement was made as well as the data upon which it was based.

---

[56] Educators, of course, have no monopoly here. Bertrand Russell reports the amusing but instructive example of Todhunter, the mathematician, who opposed the establishing of the first experimental laboratory at Cambridge, because he thought it was unnecessary for students to see experiments performed, since the results could be vouched for by their teachers, all of whom were men of the highest character, including many who were clergymen in the Church of England!

[57] Someone has suggested that theories often continue to live long after their brains have been knocked out.

The third method of arriving at truth is that of Research. This method is comparatively recent in the history of man. The prestige of the orator and rhetorician in ancient Greece and Rome and the authority of Aristotle in the Middle Ages testify to the newness of the method of research. And it is more recent in education and the other social sciences than in the physical and biological sciences. Its appeal is to the intellect and is based upon the facts in the case. It is the distinctive method of science and may be regarded as the only final method of settling an educational issue. A single illustration will indicate its superiority over the earlier methods used.

A practical problem in education is to determine the proper amount of time to be devoted to each subject taught. Educators have usually assumed that the results obtained are directly proportional to the amount of time expended. In fact, the thing seemed self-evident. In the closing years of the last century, however, an inquisitive American by the name of Rice undertook, apparently for the first time, to subject the question to scientific study. The subject chosen was spelling, and the procedure was extremely simple and direct. A uniform, although not standardized, spelling test was administered to schools in various parts of the country. Afterward the results, involving about 100,000 cases, were tabulated according to the amount of time devoted to spelling in the school program. Contrary to the usual assumption, Rice found little or no relation between the results obtained and the time expended.[58] Equally good spelling achievement was found in schools where a period of ten or fifteen minutes was devoted to the subject as in those where a period three or four times as long was allowed. Although considerable skepticism was manifested toward the Rice inquiry at the beginning, the evidence was so convincing as to compel assent. Today hardly any school allots more than fifteen minutes a day to spelling in the school program. The solution of practical problems in education by the method of research had thus made a promising beginning.

Nearly half a century later Tyler summarized the educational situation as follows: [59]

---

[58] Later studies have found a similar lack of relationship in other subjects. See Merrill B. Eaton, "A Survey of the Achievement in Social Studies of 10,200 Sixth Grade Pupils in 464 Schools in Indiana," *Bulletin of the School of Education, Indiana University*, 20: 53, May, 1944.

[59] *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II*, page 349. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1938.

> The proceedings of educational associations during the latter part of the nineteenth century indicate clearly an attempt to settle teaching problems by argument, by impassioned pleas, or by consensus. The achievement-testing movement provided a new tool by which educational problems could be studied systematically in terms of more objective evidence regarding the effects produced in pupils. The hope that problems could be settled by reference to fact rather than subjective impression or emotionally colored opinions has probably been the strongest influence of the achievement-testing movement in the past forty-one years.

A splendid discussion of the present status of research in education has been made by Good, Barr, and Scates.[60] These authors classify research methods in education under four headings, historical, normative-survey, experimental, and other methods. Of these four methods only the historical is not dependent upon measurement in some form, and even this method is likely to make use of numerical data.

The function of measurement in instruction and in school administration. The foregoing discussion has been primarily concerned with the role of measurement in educational research, or in education considered as a science. But education is an art as well as a science. It has its practical as well as its theoretical aspects. It is the primary purpose of this book to consider such immediately practical problems as the actual administration of schools and the instruction of pupils in these schools. Later chapters will consider the following topics:

A. Measurement in Instruction.
  1. Motivation.
  2. Practice.
  3. Diagnosis.
  4. School marks.
B. Measurement in School Administration.
  1. Classification and promotion.
  2. Guidance.
  3. Evaluation.
  4. Public relations.

There is some overlapping among these divisions, each of which could be subdivided still further. But the organization is a convenient one, even if somewhat arbitrary.

It must be admitted at the outset that up to the present time research, while not entirely lacking, is by no means sufficient to prove conclusively that measurement really serves the above practical functions, or even that schools cannot conceivably be operated effectively without tests and examinations of any kind.[61] It is to be regretted, however, that the case for measurement in education must for the present rest largely upon the testimony of experienced teachers and school administrators, and the argumentative ability of persons enjoying the highest reputation in the field. This is one of the many points in education where further research is sadly needed. It is certainly most disquieting to find such outstanding champions of measurement in education as McCall,[62] Symonds,[63] and Ruch [64] unblushingly defending their cause by argument rather than by experimental evidence. It is an anomalous situation indeed when science attempts to establish its value by an appeal to philosophy!

---

[60] Carter V. Good, A. S. Barr, and Douglas E. Scates, *op. cit.*, 882 pages.

Examples of existing experimental evidence as to the value of measurement in instruction on the college level are the following: [65]

> Schutte found that normal school students who expected final examinations did significantly better than those who did not. Kulp found weekly tests increased the amount learned in educational sociology by about 17 per cent. Turney found that educational psychology students who took twelve short tests did about 20 per cent better than others who took only the mid-term and final examinations. Jones found that psychology students who took a five-minute test after each lecture retained after eight weeks approximately twice as much as those who did not. Keys found that the same tests administered in the form of weekly rather than monthly examinations in educational psychology gave an immediate superiority of 12 per cent.

It must be recognized that these are but straws, which may or may not tell which way the wind blows. With a few notable exceptions the instructional values of tests and examinations in the elementary and high school have received much less attention. Even on the college level the evidence leaves much to be desired. The number of cases has usually been small, the period of time short, and the measurement of results limited to informational learning at the end of the course. Little or no account has been taken of the permanent effects of the learning or the effect of the testing on the point of view of the student, or his ability to put his knowledge to the solution of actual problems. It is evident, therefore, that in spite of the universal use of measurement and evaluation in some form, educators have taken its values largely on faith. And it is most surprising that educational research, which is dependent for its very existence upon measurement, has thus far given so little attention to the value of measurement in the practical affairs of the school, in spite of the fact that this value gives educational research its excuse for being.

---

[61] It must be emphasized, however, that while the experimental evidence for the practical value of measurement is meager and inconclusive, objective support for the view that measurement is useless or harmful is altogether lacking.

[62] William A. McCall, *op. cit.*, Chapter I.

[63] Percival M. Symonds, *Measurement in Secondary Education,* Chapter I. New York: The Macmillan Company, 1927.

[64] G. M. Ruch, *The Objective or New-Type Examination,* Chapter I. Chicago: Scott, Foresman & Company, 1929.

[65] These studies and other related studies will be discussed more fully in Chapter XI and Chapter XII.

Types of measurement in education. The various types of measurement employed in education may be classified on different bases and from different points of view. The following appears to be a reasonably satisfactory classification of the instruments of measurement employed in the ordinary school for distinctly educational purposes:

A. Oral.

B. Written.
  1. Informal (nonstandardized).
    a. Traditional (essay type).
    b. Objective (new-type).
  2. Formal (standardized).
    a. Achievement.
      (1) General (survey).
      (2) Specific (diagnostic, practice, etc.).
    b. Intelligence.
      (1) General (individual and group).
      (2) Specific (aptitude or prognosis).
    c. Character and Personality.

The distinction between the major categories, oral and written, is obvious. The distinction between informal and formal written tests is also easy to make. A formal test often begins as an informal test, which is later subjected to experimental trial and revision, only the best items surviving the process. Formal tests also have carefully worded instructions both for administering and scoring and, usually, norms for interpreting the results.

The distinctions among tests of achievement, intelligence, and character and personality are not so clear-cut, however. By the term *achievement tests* is meant tests of academic achievement, such as arithmetic or algebra; they are distinguished from character and personality tests, which are also largely tests of achievement, but mainly of a different sort. Intelligence tests, theoretically at least, are measures of learning capacity, whereas achievement tests are measures of learning itself. In other words, intelligence tests attempt to measure educability, while achievement tests attempt to measure education. The writer has followed the usual practice of recognizing tests of achievement and intelligence as co-ordinate with tests of character and personality. Strictly speaking, however, achievement and intelligence are merely aspects of personality, which is a term used by psychologists to include every trait which differentiates one individual from another. In a sense, then, every test is a test of personality; and many aspects of personality cannot be measured by tests at all, but are evaluated by means of rating scales, questionnaires, interviews, controlled observation, and the like.

Tests are subdivided into general and specific on the basis of scope. They may also be further subdivided into individual and group tests on the basis of method of administration, and into verbal and nonverbal or performance tests on the basis of content. A distinction is often made, although not always observed in practice, between a test and a scale. A test consists of a series of questions to be answered or exercises of some sort to be done, and a pupil's performance is the number of these he is able to do in the time allotted. Strictly speaking, a scale consists of a series of specimens, such as handwriting, for example, arranged in order of merit, and the pupil's performance is judged by comparing it with the standard specimens. Most standardized instruments of measurement are really tests rather than scales. The test items are frequently arranged in order of difficulty, however, in which case the term *scaled test* is sometimes used to distinguish such tests from those in which the items are not so arranged.

Of course, many other types of measurement are employed in schools. Examples of these are chronological age, height, weight, temperature, and time, but these can hardly be classified as strictly educational. It cannot be too strongly emphasized that measurement is not limited to tests and examinations, and certainly not to standardized tests. There are also numerous rating scales and check lists for playgrounds, buildings and equipment, and so forth, whose use is largely restricted to the specialist. These have been omitted in the interest of brevity.

It must be recognized that recent tendencies in education have enlarged its scope and increased its complexity, and have thereby added to the difficulties of teaching and administration. But nowhere have these difficulties been more apparent than in the problem of measurement. The need for proper evaluation is as great in the modern school as ever before, but the difficulties of providing for it are vastly greater. For, as Saucier points out, "an instrument of measurement may meet all the criteria for measuring a reactionary, undemocratic conception of education but at the same time be valueless for measuring the major results of a progressive, democratic theory of education."[66] This means that as the schools improve, so must the tools and techniques of measurement and evaluation.

---

[66] W. A. Saucier, *Introduction to Modern Views of Education,* pages 368–369. Boston: Ginn and Company, 1937.

SELECTED REFERENCES FOR FURTHER READING

Brubacher, John S., *Modern Philosophies of Education.* New York: McGraw-Hill Book Company, Inc., 1939. Chapter I.

Dampier, Sir William Cecil, *A History of Science* (Third Edition). New York: The Macmillan Company, 1942. 574 pages.

Dewey, John, *Logic, The Theory of Inquiry.* New York: Henry Holt & Company, 1938. Chapter XI.

———, *The Sources of a Science of Education.* New York: Liveright Publishing Corporation, 1929. 77 pages.

Doughton, Isaac, *Modern Public Education, Its Philosophy and Background.* New York: D. Appleton-Century Company, 1935. Chapters I, VI, VII, VIII, and IX.

Freeman, Frank N., and others, "The Scientific Movement in Education," *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II.* Bloomington, Illinois: Public School Publishing Company, 1938. Chapters XXXVIII and XXXIX.

Good, Carter V., Barr, A. S., and Scates, D. E., *The Methodology of Educational Research.* New York: D. Appleton-Century Company, 1936. Chapter I.

Jeans, Sir James, *Physics and Philosophy.* New York: The Macmillan Company, 1943. 222 pages.

McCall, William A., *Measurement.* New York: The Macmillan Company, 1939. Chapter I.

Mayer, Joseph R., *The Seven Seals of Science.* New York: D. Appleton-Century Company, 1927. 444 pages.

Russell, Bertrand, *The Scientific Outlook.* New York: W. W. Norton & Company, Inc., 1931. 277 pages.

Westaway, F. W., *Scientific Method: Its Philosophical Basis and its Mode of Application* (Fifth Edition). New York: Hillman-Curl, Inc., 1937. 588 pages.

Whitehead, Alfred North, *Science and the Modern World.* New York: The Macmillan Company, 1925. 304 pages.

Whitney, Frederick Lamson, *The Elements of Research.* New York: Prentice-Hall, Inc., 1937. Chapter I.

CHAPTER II

The Historical Development of Measurement in Education

A. Introduction

Tests and measurements of one kind or another have played a far more prominent role in human history than is generally recognized. Nor has their use by any means been confined to the schools. In fact, among the earliest records of the use of various testing devices are those found in the Bible, although they generally have no direct reference to education. One illustration[1] will suffice:

> And the Gileadites took the passages of Jordan before the Ephraimites: and it was *so*, that when those Ephraimites which were escaped said, Let me go over; that the men of Gilead said unto him, *Art* thou an Ephraimite? If he said, Nay; then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce *it* right. Then they took him, and slew him at the passages of Jordan: and there fell at that time of the Ephraimites forty and two thousand.

Attention is called to the fact that here is indeed a "final examination" and in a field other than education. Doubtless measurement experts of the present time would point out that, in spite of a rather high degree of objectivity, there were certain dubious features: it was oral, it was very short, and the mortality rate was excessively high!

A sociologist[2] attributes the remarkable stability of the Chinese civilization, the oldest culture of any modern nation, to five factors, one of which is her highly organized examination system. It began informally in 225 B.C., and became a definite civil service examination system in 29 B.C. The system, described as being thoroughly democratic, ruthless, invariable, and orthodox, has had profound effects, some good and some bad, not only upon the educational system of China, but also upon her whole civilization. On the one hand, it has preserved unity by keeping uniform throughout the empire the written language, literature, and traditions of the Chinese nation, and has helped to maintain political stability by keeping open to every citizen the door to prestige and power. On the other hand, it has often produced more graduates than could be given positions, has offered little assurance that the successful candidates possessed the qualities necessary for good officials, has sometimes resulted in corruption in the conduct of the examinations, and has in some degree stifled progress.

---

[1] *Judges,* 12: 5–6 (King James Version).

[2] Paul F. Cressey, "The Influence of the Literary Examination System on the Development of Chinese Civilization," *American Journal of Sociology,* 35: 250–262, September, 1929.

Some kind of measurement or evaluation seems inevitable in education. It seems inherently an essential part of the teaching process.[3] The situation has been well expressed [4] by Dean McConn as follows:

> As far back as we have any record of school routines, teachers have always examined or tested, as well as taught. But our attitudes towards these two functions have been, historically, quite different. We have long understood that teaching is a highly skilled business, a profession, calling for special aptitudes and extensive preparation, and that both its techniques and its objectives are worthy of the most careful investigation. But examining or testing we have taken for granted as something that anybody could do any time, quite casually, for any purpose he might happen to think of. It is only yesterday that it occurred to most of us that there might be skilled techniques of testing or that the uses we were accustomed to make of our tests and examinations might be open to question.
>
> Every teacher or administrator of more than twenty years' service will recall with me that Age of Innocence when a "test" regularly consisted of ten questions, sometimes concocted impromptu as we wrote them on the blackboard, each weighted, by our arbitrary personal fiat, with a value of 10 on a scale of 100, and when the perfectly simple purpose of any "test" was to "pass" or "flunk" the testees. We knew no qualms in those days about reliabilities or validities or comparability, and the sigma lay as far in the future (for us teachers) as television.
>
> It was only after the World War that this primal innocence was disturbed by the coming—into the consciousness of teachers generally, as distinguished from the psychologists—of what many of us still think of as "the new tests." A bewildering series of strange inventions: intelligence tests first, and then objective achievement tests, and aptitude tests, and interest tests, and personality inventories and ratings. Nearly all of them appallingly elaborate; and alleged to have been most laboriously prepared, with every item studied and checked, and to have been tried out on hundreds or thousands of students, and then re-studied and re-checked by mysterious statistical methods.

---

[3] In this connection the recent experience of Russia is instructive. In 1917 the Soviet Government, wishing to achieve as complete a transformation of education as of government, did away with all forms of examinations and school marks. After fifteen years' experience, however, the Central Committee of the Communist Party declared the plan ineffective and undesirable, and recommended the reintroduction of a rigid system of examinations and marks, a policy which has since been adopted. See *School and Society,* 40: 477, October 13, 1934; 42: 836–837, December 14, 1935; and 50: 25–26, July 1, 1939.

[4] Max McConn, "Examinations Old and New: Their Uses and Abuses," *Educational Record,* 16: 375, October, 1935.

The foregoing picturesque statement is accurate enough for "teachers generally," as the author intended, but is far from true of certain outstanding leaders in the profession. Horace Mann,[5] for example, almost one hundred years ago, had a remarkable conception both of the importance of examinations and of the limitations of the forms then in existence. His penetrating analysis of the weaknesses of the oral examinations then in vogue, and of the superiority of written examinations, could hardly be improved upon by the modern specialist in measurement. Mann showed clearly the points where the oral examinations were lacking, in the technical language of today, in validity, reliability, and usability.[6]

Another American educator who understood both the value and the limitations of examinations was Emerson E. White, widely known as a writer and school administrator. More than half a century ago he wrote: "It may be stated as a general fact that school instruction and study are never much wider or better than the tests by which they are measured." [7] In the same volume [8] the author enumerates several "special advantages" of the written test:

> It is more impartial than the oral test, since it gives all the pupils the same tests and an equal opportunity to meet them; its results are more tangible and reliable; it discloses more accurately the comparative progress of the different pupils, information of value to the teacher; it reveals more clearly defects in teaching and study, and thus assists in their correction; it emphasizes more distinctly the importance of accuracy and fullness in the expression of knowledge; it reveals more fully than the ordinary language exercise the ability of the pupil to write correctly when his attention is directed to the thought or subject-matter; it is at least an equal test of the thought-power or intelligence of pupils, since this result, in both methods, is dependent upon the nature of the tests; and, lastly, the certainty of the coming written test affords a healthy stimulus to pupils, increasing their attention to instruction, and their efforts to master the subjects taught.

These views of Mann and White appear surprisingly modern and show how far the practice of the rank and file is likely to fall behind the theory of the pioneer thinker. It is doubtful if any single sentence in recent educational literature states the superiority of the written over the oral examination more completely or more forcefully than the one just quoted from White. In fact, the modern specialist in measurement would probably accept the above indictment of oral tests *in toto*. He would, of course, wish to discount somewhat the values so enthusiastically proclaimed for ordinary written examinations, and he would point out that many of the limitations of the oral tests so forcefully stated also hold in some degree for the written tests, and in addition that the latter have some special limitations of their own not then recognized. But that is another story to be told later.

---

[5] Otis W. Caldwell and Stuart A. Courtis, *Then and Now in Education: 1845–1923,* pages 37–41. Yonkers: World Book Company, 1923.

[6] These terms will be explained in the next chapter.

[7] Emerson E. White, *The Elements of Pedagogy,* page 148. New York: American Book Company, 1886.

[8] *Ibid.,* pages 197–198.

B. The History of Intelligence Tests

In Jevons' *The Principles of Science,* published in 1874, occurred this significant statement: [9]

> As physical science advances, it becomes more and more accurately quantitative. Questions of simple logical fact after a time resolve themselves into questions of degree, time, distance, or weight. Forces hardly suspected to exist by one generation, are clearly recognised by the next, and precisely measured by the third generation. But one condition of this rapid advance is the invention of suitable instruments of measurement. . . . Accordingly the introduction of a new instrument often forms an epoch in the history of science.

While the foregoing statement was intended as a history of the past development of physical science, it is also a remarkably accurate prophecy of the future development of measurement in psychology, which Jevons appears to have foreseen, as indicated by his reference to the "fact that man in his economical, sanitary, intellectual, aesthetic, or moral relations may become the subject of exact sciences, the highest and most useful of all sciences." [10] This statement is all the more remarkable when one considers that it was made five years before Wundt established the first psychological laboratory and Galton began publishing his most important studies of individual differences, while both Binet and Cattell were lads in their teens, and before either Thorndike or Terman had been born. But it was over a quarter of a century before any very definite progress was made toward fulfilling the prophecy. Then there followed rapid progress in that direction along several lines. That story will now briefly be told.

Germany and experimental psychology. An important event in the history of psychology was the establishment of the first experimental laboratory in psychology by Wilhelm Wundt at Leipzig in 1879. He was, however, primarily interested in the analysis of consciousness into elements in a manner analogous to that employed in atomic chemistry. His sole interest in measurement appeared to be confined to reaction times, and he was distinctly unsympathetic to the problem of individual differences; but he did influence considerably the course of psychology, especially the work of other German psychologists, such as Kraeplin, Ebbinghaus, and Meumann, who introduced many forms of separate tests, which were borrowed by later investigators in constructing their scales for measuring general intelligence. Of the test forms, the completion test of Ebbinghaus was doubtless the most important.

---

[9] W. Stanley Jevons, *The Principles of Science,* Book III, page 313. New York: The Macmillan Company, 1874.

[10] *Ibid.,* page 386.

Another important idea, suggested in 1912 by Stern, was that of representing intelligence as the ratio of mental age to chronological age. This concept, for which Stern suggested the term "mental quotient," was later adopted by Terman as the familiar IQ.

England and statistical methods. The distinctive contribution of the English to the measurement of intelligence has been that of statistical methods as a tool for the analysis of test results. Sir Francis Galton, one of the most brilliant and versatile men of the nineteenth century, was the first to treat seriously the problem of individual differences in psychology, particularly in the realm of sensory discrimination, although Weber, Fechner, Helmholtz, and others had given slight attention to it in what is often termed psychophysics. In 1883 Galton outlined a method for studying free association by quantitative methods. But his most notable contribution was in statistical analysis, where he suggested among other things a graphical method of representing correlations.[11] Karl Pearson, a pupil of Galton, and Charles E. Spearman still further advanced the science of statistics. Spearman developed his well-known two-factor theory of intelligence on the basis of statistical analysis. Cyril Burt, who has been a leader in introducing and adopting Binet's work in Great Britain, was in 1913 officially appointed school psychologist, possibly the first person in the world to occupy that position.

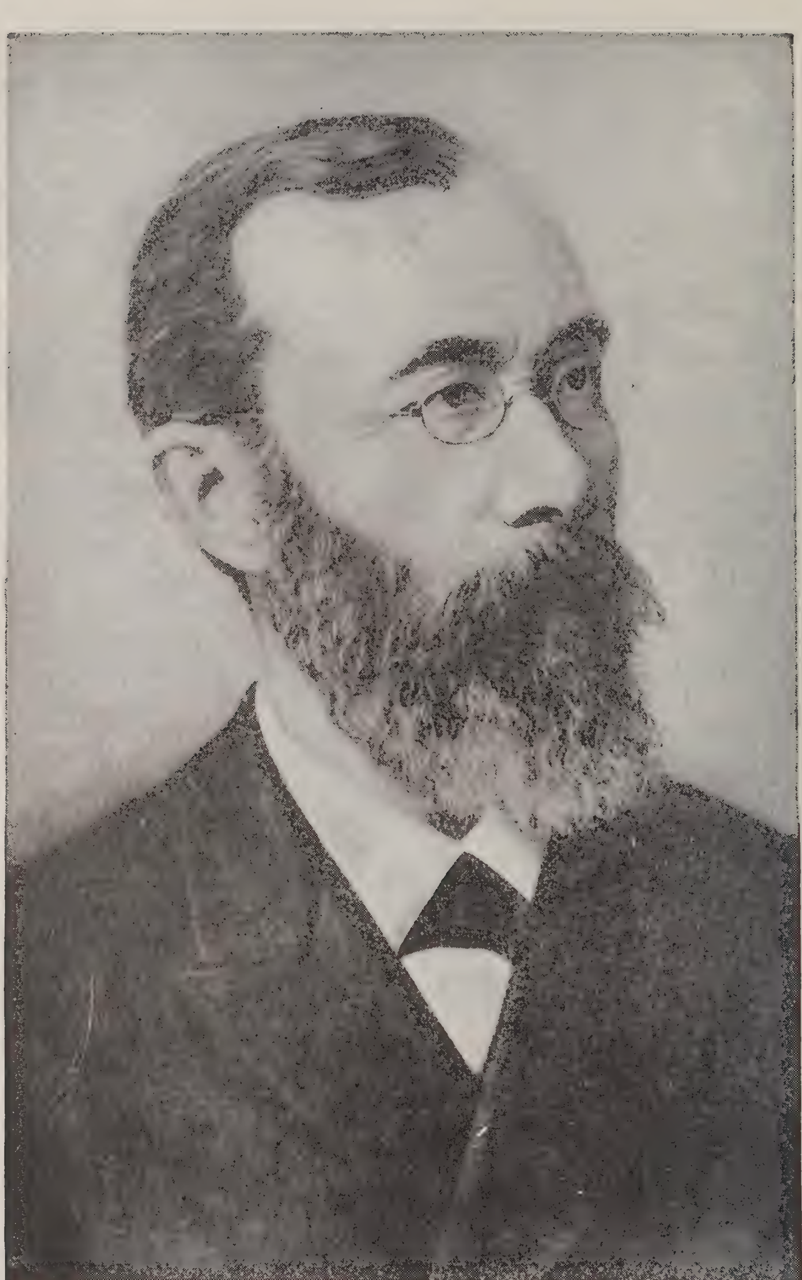
France and abnormal psychology. The French have always been leaders in abnormal psychology. Consequently, they approached the problem of measuring intelligence from the standpoint of the classification and treatment of the mentally defective. This brings us to the most important name in the history of intelligence testing, Alfred Binet.

It would be hard to find a man who better illustrates Jevons' description of the methods of the genius than Binet. Jevons says: [12]

---

[11] David G. Ryans, "Francis Galton's Statistical Contributions," *School and Society,* 48: 312–316, September 3, 1938.

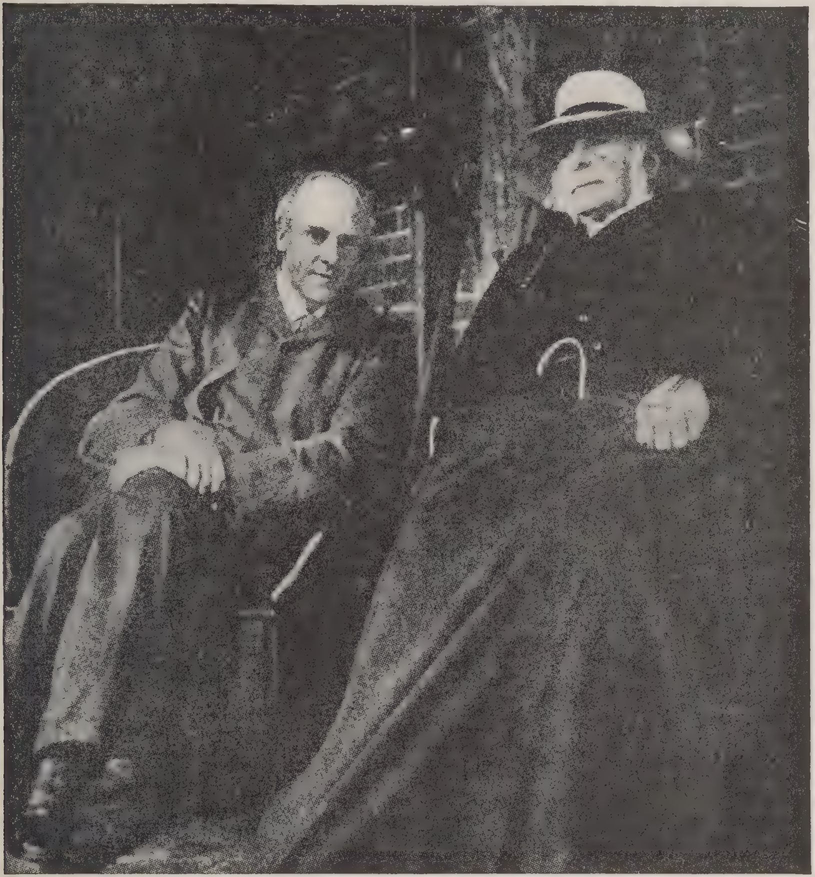
[12] W. Stanley Jevons, *op. cit.,* Book IV, page 221.



Reproduced by permission of the C. H. Stoelting Co., Chicago; courtesy Open Court Publishing Co.

Wilhelm Wundt (1832–1920)

> It would be a complete error to suppose that the great discoverer is one who seizes at once unerringly upon the truth, or has any special method of divining it. In all probability the errors of the great mind far exceed in number those of the less vigorous one. Fertility of imagination and abundance of guesses at truth are among the first requisites of discovery.



Courtesy of Professor Helen Walker.

Karl Pearson (1857–1936) and Sir Francis Galton (1822–1911)

This reads as if it were designed specifically to describe Binet, and yet it appeared twenty years before Binet founded *L'Année Psychologique* and thirty years before his first scale for measuring intelligence. He first studied law, then medicine, and afterward worked in a biological laboratory. Later he turned psychologist, first of the arm-chair variety, and finally ended as an experimentalist. Furthermore, in an effort to devise a suitable method of measuring intelligence he tried out various head measurements, physiognomy, graphology, and palmistry, before hitting upon the correct approach. Binet never seemed to be quite sure what he meant by "intelligence," what he was trying to measure, for he changed repeatedly his definitions. It is clear, therefore, that he did not hit "at once unerringly upon the truth," and that he did possess to a marked degree "fertility of imagination and abundance of guesses." Such errors as he made, and they were numerous, were not the unintelligent ones of blind trial and error, but rather the intelligent errors of judgment, made by acting upon the course which seemed most promising from a survey of the best available facts at hand.

It is doubtless true, as Boring suggests: [13]

> At close view, the course of science seems discontinuous; all at once a "genius" makes a discovery or formulates a theory, and productive research follows on immediately. At the greater range of historical perspective, the course of science seems to be continuous, and the "genius" appears as an opportunist who takes advantage of the preparation of the times.

Opinions will differ regarding the appropriateness of the word "opportunist" in Binet's case, but there can be no doubt that he did take "advantage of the preparation of the times." Both for his ideas and actual test materials he drew freely from others, notably his fellow countryman, Blin, and his contemporaries in Germany. Nevertheless, Binet did something the others had not done; he began where they left off and continued with a definite contribution both to the theory and practice of testing. On the theory side he enlarged the prevailing concept of intelligence, introducing such ideas as those of judgment, adaptation, and self-criticism. Terman [14] argues that Binet's outstanding contribution to psychometrics was his abandonment of any attempt to measure "intellectual faculties as such." To practice he contributed a technique of scale construction and a finished scale consisting of test situations selected according to predetermined criteria and standardized. The date 1905 is important, therefore, because it marked the appearance of the first scale for the measurement of intelligence, which, crude as it was, has served as the pattern for all subsequent tests and scales the world over. The 1908 revision was a definite improvement, and is especially notable for the introduction of the *mental age* concept. Further experimental work resulted in the scale of

---

[13] Edwin G. Boring, *A History of Experimental Psychology,* page 452. New York: The Century Company, 1929. Used by permission of D. Appleton-Century Company.

[14] Quinn McNemar, *The Revision of the Stanford-Binet Scale,* page 6. Boston: Houghton Mifflin Company, 1942.

Reproduced by permission of the C. H. Stoelting Co., Chicago; courtesy Open Court Publishing Co.

Alfred Binet (1857–1911)

1911, the year of Binet's death. Binet's successors in France have continued to interpret the results of intelligence tests in terms of MA and without recourse to the IQ.

America and applied psychology. The scene now shifts to America, where the outstanding name is J. McKeen Cattell, who was a pioneer along many lines and a promoter of the first rank. More than anyone else, Cattell was responsible for giving to American psychology its practical bent, for with him the practical took precedence over the philosophical. As early as 1885 he began to publish important articles on reaction times and individual differences. It was Cattell[15] who in 1890 suggested the term "mental tests," which was to become a sort of trade-mark for the whole measurement movement. But Cattell was too close to Wundt's laboratory to escape altogether the views of its master. Cattell, therefore, just as did Galton, confined his tests largely to the simpler mental processes, such as sensory discrimination, where individual differences are least, rather than to the higher mental processes, where they are greatest. In other words, both Galton and Cattell attempted to measure intelligence, but with the wrong tools. Very little attention was given either to reliability or to validity. Consequently, in 1901, when Wissler[16] published his analysis of Cattell's tests used with college students, in which he employed for the first time the Pearson correlation technique to test scores, and in which he found little more than chance relationship either among the tests themselves or between the tests and college work, a considerable damper was thrown over the enthusiasm of American testers which was not lifted till after Binet had published his 1905 scale. Nevertheless Cattell's influence upon measurement, through both his journals and his students, notably Thorndike, has been great.

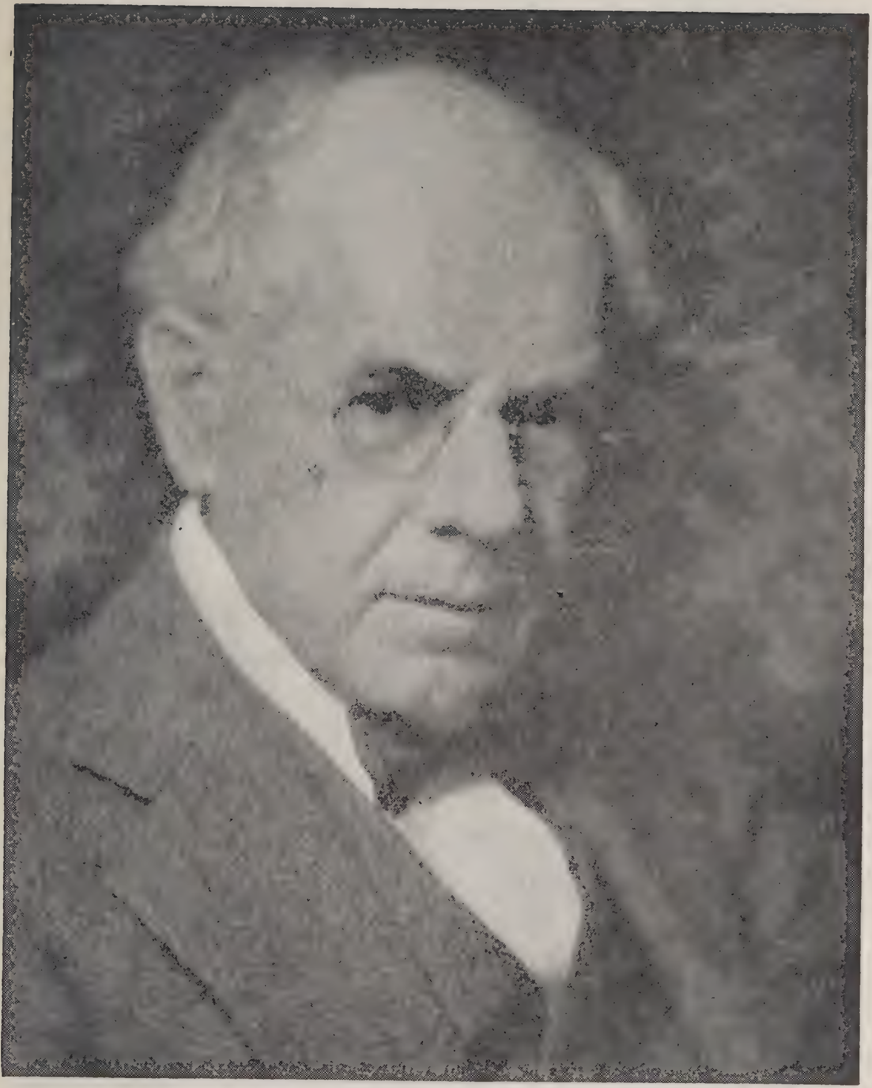
Goddard was the first American psychologist to recognize the value of Binet's 1905 scale, which he translated and with minor adaptations tried out at Vineland. In 1910 appeared two translations of the 1908 scale, one by Goddard and the other by Huey. In 1912 Kuhlmann published his first revision of the Binet scale, extending it downward to the age of three months, instead of three years, which was Binet's lower limit.

It remained for Terman of Stanford University to provide the first thoroughgoing revision, carefully adapted to and standardized for use with American children, normal as well as subnormal. Terman's scale, known as the Stanford Revision or Stanford-Binet, appeared in 1916, together with a most complete manual, *The*

---

[15] J. McK. Cattell, "Mental Tests and Measurements," *Mind,* 15: 373–380, 1890.

[16] Clark Wissler, "The Correlation of Mental and Physical Tests," *Psychological Review,* Monograph Supplement, Vol. VIII, No. 16, 1901.



James McKeen Cattell (1860–1944)

Measurement of Intelligence.[17] This revision has been criticized on the ground that it was standardized entirely on school children, which may result in somewhat of a handicap for those of poor academic background, and that it did not produce a sufficient "scatter" in the distribution of IQ's, particularly at the higher ages. It has also been criticized on the ground that its norms were based exclusively on the children of one state, California, which may not be truly representative of the United States as a whole. Nevertheless, the Stanford Revision was, for more than two decades, the most widely used and most highly regarded individual intelligence test in existence. In 1937 a thorough revision of the Stanford-Binet appeared.[18] This second revision corrected most of the weaknesses of the first revision, which it has largely supplanted.

Two other distinctly American developments, both aiming to make intelligence tests more practical, remain to be discussed These early tests had two practical disadvantages which militated against their wide use. One of these was that the tests were highly *verbal;* that is, their successful administration required that the subject taking the test understand the English language. The other was that the tests were *individual;* that is, only one person could be examined at a time. Reasonably satisfactory solutions came to both of these problems in the year 1917; and this leads to an interesting story.

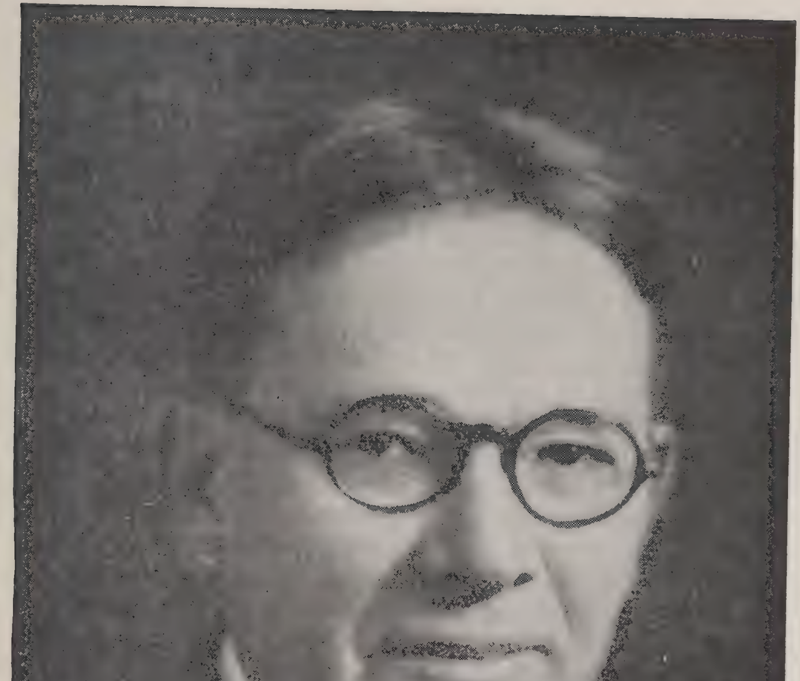
Intelligence tests, children of necessity. One cannot but be impressed with the curious role of necessity in the development of intelligence testing both in Europe and in America. Although it may be true, as Thorndike suggests, that necessity is not the true mother of invention, she is, often at least, the stern, relentless stepmother. Two instances in Europe and two in America will suffice to make this clear.

In 1897 Ebbinghaus was appointed on a commission to investigate the problem of fatigue in the schools of Breslau. As there were in existence no appropriate tests, Ebbinghaus set about to devise them. The result was the "completion tests," which have since been widely accepted as measures of general intelligence. Seven years later the Minister of Public Instruction in Paris became concerned about the high percentage of failure in the Paris schools and appointed Binet on a commission to determine those who were so mentally unfit as to necessitate instruction in special classes. Binet, too, found available measuring instruments inadequate for

---

[17] Lewis M. Terman, *The Measurement of Intelligence,* 362 pages. Boston: Houghton Mifflin Company, 1916.

[18] Lewis M. Terman and Maud A. Merrill, *Measuring Intelligence,* 461 pages. Boston: Houghton Mifflin Company, 1937.



Lewis M. Terman (1877-   )

the purpose.[19] Out of this difficulty emerged the 1905 scale already referred to, the first successful instrument for measuring intelligence according to modern conceptions. But the original Binet scales and their early revisions, both in Europe and America, possessed the two limitations mentioned above: namely, they were highly linguistic and they were individual scales. Soon American ingenuity was to offer solutions to both problems.

Pintner and Paterson, finding the Stanford-Binet unsatisfactory for deaf children, met the first difficulty by assembling a series of fifteen tests of manipulation or performance, such as the form board already used by Seguin, Healy, and others. This combination, which appeared in 1917, is known as the Pintner-Paterson Performance Scale. That same year the United States found herself in the World War, faced with the urgent necessity of training a large citizen army with an insufficient supply of commissioned and noncommissioned officers. In this emergency the American Psychological Association placed its services at the disposal of the War Department. The existing individual intelligence tests were not only entirely unsuited for use with illiterate and foreign-speaking recruits, but they were also much too slow. To meet this need, a committee of psychologists, utilizing largely the as yet unpublished work of Otis, prepared the Army Alpha,[20] the first of a long succession of group tests destined to receive wide use. The second difficulty of the early tests had now been solved, for a group test can be administered to a hundred or more in the time formerly required for measuring one.

It should be noted, however, that the Pintner-Paterson Performance Scale and the Army Alpha group tests each solved but one difficulty at a time. In fact, as a rule, group tests of the Army Alpha type are even more verbal than the individual tests had been. Figure 1 shows a sample page from the Army Alpha. The early performance scales, on the other hand, were nonverbal, but could be administered to only one person at a time. The Army Beta, designed for illiterate and foreign-speaking soldiers, was the first test to combine the group and performance ideas. Figure 2 shows the picture completion test of the Army Beta. It also appeared in 1917. Since that time several group tests, largely or primarily of the performance type, have been designed specifically for use with young children just entering school. There can be little doubt that World War I gave a decided impetus to the measurement movement in America. It is probable that World War II will have a similar effect.[21]

---

[19] Binet confessed himself unable to distinguish an idiot, who was described by existing standards as having a "gleam" of intelligence and an attention which was "fugitive," from an imbecile, who was described by these standards as having a "very incomplete degree" of intelligence and an attention which was "fleeting." Verily, here indeed was a distinction without a difference.

[20] Although the Army Alpha had antecedents developed over the preceding 30 years, none of these earlier tests can be said to have passed beyond the experimental stage.

Tests of specific intelligence, or aptitude. All of the tests so far described have been for the measurement of *general* intelligence. There has also been some activity in the development of tests of *specific* intelligence, or capacity in a restricted area, such as music or mechanics, or in a specific school subject, such as algebra or Latin. America has also had the lead in the development of these tests, often called aptitude or prognosis tests. One of the earliest and in many ways one of the best known of these tests is the Seashore Test of Musical Talent, which appeared in 1915. Three years later appeared the Stenquist Test of General Mechanical Ability. In 1918, also, Rogers published a test of mathematical ability, which, although hardly an aptitude test in the modern sense, introduced the idea which other authors have followed up by aptitude tests in the special branches of mathematics, such as algebra and geometry. A somewhat different type of test on the college level is illustrated by the Iowa Placement Examinations which appeared in 1924. A recent and promising type of test, of which there are several examples, is that of reading readiness, to be used to determine a child's fitness for the work of the first grade. There is evidence that the development of the future is likely to be along the line of tests for *specific* intelligence, or aptitude in a restricted area, rather than tests of *general* intelligence, which aim to cover the whole range of human capacity at one shot. The test maker, as well as the bird hunter, may aim at too large a target. Dunlap puts the case well: [22] "The more 'general' the intelligence test, the less its value. By increasing the specificity . . . we add to its value. Charles Dudley Warner once shot a bear by 'aiming at it generally,' but it is a poor method." Thurstone's attempt to devise tests of what he terms the seven "primary mental abilities" is a move in this direction, although these tests have not yet been developed sufficiently to demonstrate marked superiority over other tests in practical schoolroom situations.

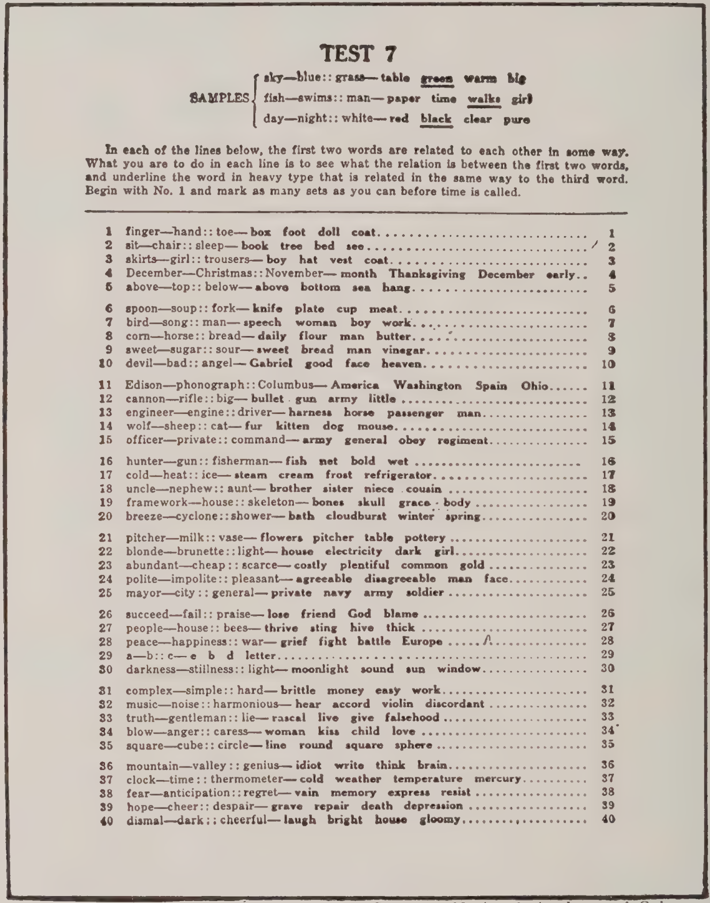
---

[21] For a discussion of the new Army General Classification Test, see: *Psychological Bulletin* 42: 760–768, December, 1945.

[22] Knight Dunlap, *Habits, Their Making and Unmaking,* page 266. New York: Liveright Publishing Corporation, 1932.

C. The History of Achievement Tests

Progress before 1918. The early history of things which have been in existence a long time is usually somewhat obscure. This is true of achievement tests, whose ancient use has already been referred to. Not only have some kinds of tests been in existence for centuries, but attention has been called to the fact that criticisms of them, both destructive and constructive, are by no means new.



TEST 7

SAMPLES
sky—blue :: grass— **table green warm big**
fish—swims :: man— **paper time walks girl**
day—night :: white— **red black clear pure**

In each of the lines below, the first two words are related to each other in some way. What you are to do in each line is to see what the relation is between the first two words, and underline the word in heavy type that is related in the same way to the third word. Begin with No. 1 and mark as many sets as you can before time is called.

1 finger—hand :: toe— **box foot doll coat** 1
2 sit—chair :: sleep— **book tree bed see** 2
3 skirts—girl :: trousers— **boy hat vest coat** 3
4 December—Christmas :: November— **month Thanksgiving December early** .. 4
5 above—top :: below— **above bottom sea hang** 5

6 spoon—soup :: fork— **knife plate cup meat** 6
7 bird—song :: man— **speech woman boy work** 7
8 corn—horse :: bread— **daily flour man butter** 8
9 sweet—sugar :: sour— **sweet bread man vinegar** 9
10 devil—bad :: angel— **Gabriel good face heaven** 10

11 Edison—phonograph :: Columbus— **America Washington Spain Ohio** 11
12 cannon—rifle :: big— **bullet gun army little** 12
13 engineer—engine :: driver— **harness horse passenger man** 13
14 wolf—sheep :: cat— **fur kitten dog mouse** 14
15 officer—private :: command— **army general obey regiment** 15

16 hunter—gun :: fisherman— **fish net bold wet** 16
17 cold—heat :: ice— **steam cream frost refrigerator** 17
18 uncle—nephew :: aunt— **brother sister niece cousin** 18
19 framework—house :: skeleton— **bones skull grace body** 19
20 breeze—cyclone :: shower— **bath cloudburst winter spring** 20

21 pitcher—milk :: vase— **flowers pitcher table pottery** 21
22 blonde—brunette :: light— **house electricity dark girl** 22
23 abundant—cheap :: scarce— **costly plentiful common gold** 23
24 polite—impolite :: pleasant— **agreeable disagreeable man face** 24
25 mayor—city :: general— **private navy army soldier** 25

26 succeed—fail :: praise— **lose friend God blame** 26
27 people—house :: bees— **thrive sting hive thick** 27
28 peace—happiness :: war— **grief fight battle Europe** 28
29 a—b :: c— **e b d letter** 29
30 darkness—stillness :: light— **moonlight sound sun window** 30

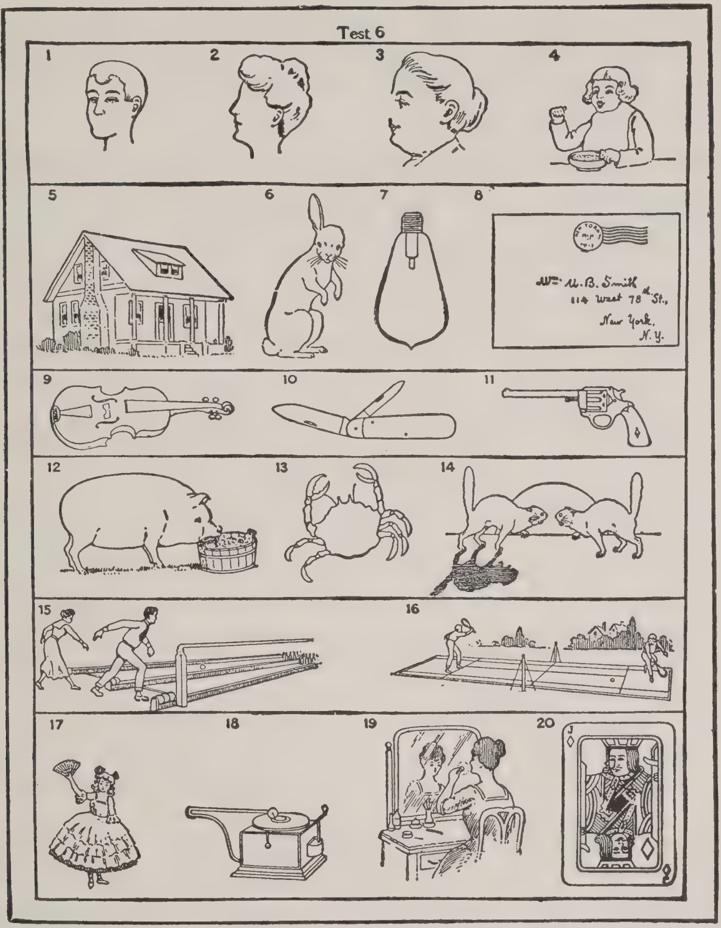
31 complex—simple :: hard— **brittle money easy work** 31
32 music—noise :: harmonious— **hear accord violin discordant** 32
33 truth—gentleman :: lie— **rascal live give falsehood** 33
34 blow—anger :: caress— **woman kiss child love** 34
35 square—cube :: circle— **line round square sphere** 35

36 mountain—valley :: genius— **idiot write think brain** 36
37 clock—time :: thermometer— **cold weather temperature mercury** 37
38 fear—anticipation :: regret— **vain memory express resist** 38
39 hope—cheer :: despair— **grave repair death depression** 39
40 dismal—dark :: cheerful— **laugh bright house gloomy** 40

Courtesy National Academy of Sciences.

Figure 1. Test 7 from the Army Alpha.

But the actual work of improving the existing instruments has always lagged far behind the theory, and actual school practice has been furthest behind of all. In spite of the marked superiority of written examinations over oral, pointed out by Horace Mann in 1845, educators did not forthwith adopt the former or improve the latter.[23] However, as early as 1864 an English schoolmaster, the Reverend George Fisher,[24] evidently realizing the subjectivity of ordinary examinations, proposed a "Scale-Book," made up of "vari-



Courtesy National Academy of Sciences.

Figure 2. Test 6 from the Army Beta.

---

[23] As Caldwell and Courtis observe: "Very few schoolmen proved to have the intelligence of Horace Mann, and the era foreseen by him did not begin to materialize until more than fifty years later." *Op. cit.*, page 8.

[24] For a good discussion of the early history of achievement tests, see: Leonard P. Ayres, "History and Present Status of Educational Measurements," *Seventeenth Yearbook of the National Society for the Study of Education, Part II,* pages 9–15. Bloomington, Illinois: Public School Publishing Company, 1918.

ous standard specimens . . . arranged in order of merit." But Ayres observes that "Mr. Fisher's efforts seem to have produced no lasting results," for which this explanation[25] is suggested:

> Progress in the scientific study of education was not possible until people could be brought to realize that human behavior was susceptible of quantitative study, and until they had statistical methods with which to carry on their investigations.

Although Ayres felt that Galton's work had largely met these two needs, he gave Dr. J. M. Rice the honor of being the "real inventor of the comparative test" in America. This was in 1894. Rice had studied in Germany and had come under the influence of experimental psychologists both at Jena and Leipzig, but the informal manner in which he carried out his famous spelling inquiry[26] on school children would probably have sent the blood pressure of Wundt up many points, had he known about it. Here again the attitude of the educational leaders toward Rice's work was anything but cordial, and "for more than ten years but little progress was made beyond the work of the pioneer himself."

Ayres makes a distinction between the "inventor" of educational measurement and the "father" of the movement. The latter distinction he awards to Dr. Edward L. Thorndike. The honor is richly merited, for no other person has touched the measurement movement at so many points or has contributed so much to it. In addition to his very influential publications on statistical methods in education and his pioneer work on intelligence tests for college entrance, either Thorndike or his students were responsible for most of the early standard tests and scales for measuring achievement. The first *test* was the Stone Arithmetic Test published in 1908, and the first *scale* was the Thorndike Handwriting Scale announced in 1909 and published the following year. The next few years saw the appearance of scales and tests in various fields. The school survey movement undoubtedly added impetus to the measurement movement, as did the appearance of certain important books and periodicals to be referred to later.

Studies in the unreliability of school marks and examinations. But there was an additional factor which served as a very strong stimulus to standard tests: *Educators discovered for the first time just how bad existing measurements were.* Beginning about 1910, several studies in rapid succession made this point convincingly clear. A distinction should be made between the limitations of school marks and the limitations of school examinations. The need for reform in college marking was forcibly brought to public attention by Max Meyer,[27] who reported on the marks collected from forty instructors for a period of five years at the University of Missouri. He found such astonishing variations as 55 per cent of *A*'s in philosophy and only 1 per cent in Chemistry III, while there were 28 per cent of failures in English II and none in Latin I. Johnson[28] found a similar condition in the University of Chicago High School. In a two-year period he found, for example, that the marks for German showed 17.1 per cent *A*'s and 8.4 per cent *F*'s, whereas the marks in English showed 6.5 per cent *A*'s and 15.5 per cent *F*'s. Such variations both at Chicago and Missouri could be most reasonably interpreted on the supposition, not that *English* is harder than foreign languages, but that English *instructors* are harder. In other words, school marks are highly subjective, the mark received often being more a function of the *personality of the instructor* than of the *performance of the student*. Further studies showed similar results elsewhere without exception. This was certainly disturbing, if not, as Thorndike suggests, actually "scandalous."[29]

---

[25] Leonard P. Ayres, *op. cit.*, page 10 quoted from original article, *via* Thorndike and Kandel.

[26] See pages 21 and 51 for further discussion of Rice's work.

But the evidence presented by a second type of study was even more damaging. Variations among the final marks in different departments might be accounted for, at least in part, by variations in the background, intelligence, and application of the students in these departments. This, at any rate, provided a comfortable loophole. But even this avenue of escape was soon to be closed. Manifestly, such factors could not be responsible for differences when several persons were marking the same student's paper, and least of all when the same person marked the same paper on two different occasions. But studies in abundance have established both conditions.

Perhaps the most striking of the early studies were those of Starch and Elliott. In one of these studies Starch and Elliott[30] used facsimile copies of the same geometry paper which were marked by 116 high-school teachers of mathematics. The values assigned ranged from 28 to 92. Manifestly, if high-school teachers cannot agree any more closely than that in mathematics, one of the most objective subjects, the situation is indeed bad.

---

[27] Max Meyer, "The Grading of Students," *Science,* 28: 243–250, August 21, 1908.

[28] Franklin W. Johnson, "A Study of High-School Grades," *School Review,* 19: 13–24, January, 1911.

[29] *Twenty-First Yearbook of the National Society for the Study of Education, Part I,* page 2.

[30] Daniel Starch and Edward C. Elliott, "Reliability of Grading Work in Mathematics," *School Review,* 21: 254–259, April, 1913.

Other studies tended but to confirm the suspicions. One of the most spectacular of these later studies was that by Falls,[31] who had 100 English teachers mark a composition by assigning it a percentage value and also indicating the school grade in which they would expect that quality of work to be done. As will be noted from Table 1, the percentage values varied from 60 to 98, and the estimated grade location, from the fifth grade to the junior year of college. As a matter of fact, the composition was the best one found by a survey committee at Gary, Indiana, a few years earlier, and was written by a high-school senior whose special interest was journalism and who was a correspondent for some of the Chicago newspapers. It is not unreasonable to suppose that many of these English teachers will never have as good a composition submitted by one of their pupils or that few of these teachers could have written a better one themselves.

TABLE 1

THE ESTIMATED GRADE-VALUE AND PERCENTAGE MARKS ASSIGNED TO AN ENGLISH COMPOSITION BY ONE HUNDRED TEACHERS (AFTER FALLS)

GRADE-VALUE	PERCENTAGE MARK								
	60–64	*65–69*	*70–74*	*75–79*	*80–84*	*85–89*	*90–94*	*95–99*	*Total*
XV								2	2
XIV									0
XIII							1	2	3
XII					1		2	3	6
XI			2			6	5	2	15
X			1	3	8	4	7	1	24
IX	1		1	1	8	4	4	3	22
VIII			2	2	2	3	4	3	16
VII				2	2	2	1		7
VI	1				1	1		1	4
V	1								1
Total	3		6	8	22	20	24	17	100

Evidence is available that examiners in other fields show variations fully as startling as those reported in public education. Ten examination papers written by applicants for licenses to practice dentistry in Kentucky were submitted for regrading to the regular examiners on the official boards of 23 other states. The results are summarized in Table 2. The papers are arranged from low to high according to the median judgment of the 24 examiners, who are designated by the letters A to X according to degree of strictness in marking. That the variations are enormous is indicated by several facts. With a minimum passing mark of 75, it will be noted that every paper was passed by at least four examiners, and failed by at least four other examiners. The most liberal examiner, A, passed them all, while the two strictest, W and X, failed them all! Seven different papers were rated by one or more examiners as the best of the ten, while two of these seven papers were rated by other examiners as the poorest of the ten. Surely such a situation can hardly be regarded as anything but chaotic.[32]

---

[31] J. D. Falls, "Research in Secondary Education," *Kentucky School Journal*, **6**: 42–46, March. 1928.

But Starch [33] also presented the problem in a different and still more unfavorable light. He found that college instructors assigned different marks when they regraded their own papers without knowledge of their former marks. In a later study Ashbaugh [34] had 49 Ohio State University seniors and graduate students, the latter with teaching experience, rate a seventh-grade arithmetic paper on a percentage basis three times, at intervals of four weeks between ratings. Some idea of the lack of consistency in scoring can be gained when it is mentioned that only one student gave the same total score on all three trials and only seven gave the same total score on any two successive trials. The mean differences between pairs of scores on successive trials were as follows: between the first and second trials, 8.1 points; between the second and third trials, 7.3 points. In studies by the writer using the same arithmetic paper, he has found variations of as much as 27 points on successive trials by the same scorer and as much as 10 points, variation on values assigned to the first problem on two successive trials approximately ninety days apart.

In a similar study, Hulten [35] found that 28 Wisconsin high-school English teachers of experience differed widely on trials at an interval of two months in the values assigned an English composition, which they thought was written by an eighth-grade pupil, but which was really part of the Hudelson Scale, at the time new and unfamiliar. He found that 15 teachers who gave passing marks the first time

---

[32] For a fuller account of this investigation, see: Leon M. Childers, "Report of the Research Committee on Examinations," *Proceedings of the Sixtieth Annual Meeting, National Association of Dental Examiners,* 60: 77–106, August, 1942.

[33] Daniel Starch, "Reliability and Distribution of Grades," *Science,* 38: 630–636, October 31, 1913.

[34] E. J. Ashbaugh, "Reducing the Variability in Teachers' Marks," *Journal of Educational Research,* 9: 185–198, March, 1924.

[35] C. E. Hulten, "The Personal Element in Teachers' Marks," *Journal of Educational Research,* 12: 49–55, June, 1925.

TABLE 2

PERCENTAGE VALUES ASSIGNED TO TEN ESSAY EXAMINATION PAPERS BY TWENTY-FOUR EXAMINERS

Percentage Values Assigned	Number of Examiners Marking Each Paper									
	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth	Ninth	Tenth
100							M			
95·99			Ⓐ						Ⓐ	FGJ
90·94					B	ⒶC	Ⓐ	ⒶFIP		ⒶM
85·89			GI	ⒶBH	ⒶCDEFM	GKO	BCEGHIN	BD	BCDEFHM	BCDEHIOQR
80·84		ⒶCEH	E	CDEF	GJ	DFHINPQR	DFJKO	CEHJKLMNOQS	GIJPQR	LN
75·79	ⒶDKL		CDJLS	GIKLMNPV	HILNORS	EMT	LPQSU	GR	KOST	KPST
70·74	EHT	BDFGKLMR	FHKOQRT	RST	KPQTV	BJL	RT	TⓍ	MNU	
65·69	IMOPRS	JNOPST	MNP	JOU	U X	SⓌ		UV		UⓌ
60·64	CGⓌ	IQ		Q		U	V		VⓌ	V
55·59	B		VⓌ	ⓌⓍ			Ⓦ	Ⓦ	Ⓧ	
50·54	F	UⓌ	B			VⓍ				Ⓧ
45·49	JQUV	V	U		Ⓦ		Ⓧ			
40·44										
35·39	N	Ⓧ	Ⓧ							
30·34										
25·29										
20·24	Ⓧ									
Median Range	66 22·79	70 35·83	73 39·96	77 55·87	78 47·90	80 51·91	82 45·100	82 59·92	83 50·90	87 62·98

would have failed the pupil the second time the paper was marked and that 11 teachers who gave failing marks the first time would have passed the pupil the second time. Studies involving English composition are especially significant, because every essay examination is a series of compositions, and when English teachers who presumably have more than ordinary skill in this field can agree neither with other teachers nor with themselves a second time, the situation is very serious. Indeed it is hard to see how it could be worse.[36]

In February, 1918, Thorndike[37] published what has proved to be probably the most influential paper that has ever appeared on educational measurements. The paper began with the well known dictum: "Whatever exists at all exists in some amount," and ended with this note of satisfaction: "Of the gains made in the past decade, we may well be proud." As he looked into the future, Thorndike saw it conditioned by a series of *if*'s:

> If those who object to quantitative thinking in education will set themselves to work to understand it; if those who criticise its presuppositions and methods will do actual experimental work to improve its general logic and detailed procedure; if those who are now at work in devising and in using means of measurement will continue their work, the next decade will bring sure gains in both theory and practice.

We shall now take a look at what really happened in the years following Thorndike's statement of possible achievements.

Progress since 1918. According to Buckingham,[38] it was in 1919 that "test-making passed from an amateur to a professional basis." A good summary of the next decade has been made by Monroe,[39] to which reference has already been made. The monograph begins with the assurance that the pioneer state of educational research is passed and that "quantity production" has been achieved. And, as is to be expected, much of the output is not up to the highest quality, when judged by modern standards. Monroe, however, detects some evidence of a growing conviction that the emphasis should be upon quality of work rather upon mere quantity.

---

[36] Nor is the situation peculiar to America. Recent studies reported in Europe reveal a situation fully as bad. In England, for example, examiners were found to reverse their judgments almost completely when they were asked to mark again the same papers they had scored a year before. See *School and Society,* 44: 364, September 19, 1936.

[37] Edward L. Thorndike, "The Nature, Purposes, and General Methods of Measurements of Educational Products," *Seventeenth Yearbook of the National Society for the Study of Education, Part II,* 1918, pages 16–24. Quoted by permission of the Society.

[38] R. B. Buckingham, "Our First Twenty-five Years," *Proceedings of the National Education Association,* 1941, page 354.

[39] Walter S. Monroe and others, *Ten Years of Educational Research, 1918–27,* 368 pages. Bureau of Educational Research Bulletin, No. 42. Urbana: University of Illinois, 1928.

Moreover, by 1927 there were already developments in new directions which represented a distinct advance. The earlier standard tests of achievement were largely of the general or *survey* type, which afforded a general all-around measure of the pupil's attainment in the subject, but which did not give the detailed information required for remedial work. The next decade saw the development of various achievement tests of a *specific* type. For example, there appeared in several fields *diagnostic* tests, whose function was to give specific information regarding the pupil's strong and weak points. Also *practice* tests were developed, especially in arithmetic, whose primary function was not so much measurement as drill. Another important development of this period was the organization of tests into batteries made up of survey tests in the more important subjects, all published in a single booklet. In 1920, two such batteries appeared, one by Pintner and the other by Monroe and Buckingham. Two years later appeared the first edition of the well known Stanford Achievement Test, which, with successive revisions, has continued to set a high standard.

There was also a rapid development of high-school tests in the major academic subjects. Even today, however, measurement in high school can hardly be said to have kept pace with that in the elementary school. There has also been some activity, but less marked, in the development of achievement tests on the college level.

There still remained at the end of the first decade of standardized tests an important need that had not been met. Confidence in the ordinary school examination had been seriously undermined by such studies as those to which reference has already been made, and as yet no suitable substitute had been found. Also, there were many fields, especially in high school and college, where there were hardly any standard tests. Even in the subjects most fully provided with such tests, they were by no means adequate to supply the needs of the classroom teacher. Furthermore, standard tests represented a considerable item of expense which school boards at that time were often reluctant to assume. The so-called *objective,* or *new-type,* test was devised to meet just this need. McCall[40] seems to have been the first to suggest this type of test, which was merely an adaptation by the classroom teacher of the form of the test items used in the standard test. Such tests were usually mimeographed, but they were not standardized. Soon they were widely and often uncritically used.

---

[40] William A. McCall, "A New Kind of School Examination," *Journal of Educational Research,* 1: 33–46, January, 1920.

Recently, Monroe[41] has given a brief summary of the measurement movement for the quarter of a century beginning in 1920. It was during these eventful years that educational measurement passed from early adolescence to early adulthood.

Improved examinations, children of necessity. Attention was called earlier in the chapter to the role of necessity in the development of intelligence tests. Much the same influence is evident in the development of improved measurement of achievement. The origin of the objective test referred to above is a case in point. Three other instances will be cited briefly.

It was customary in the early days for the school committees in Massachusetts to give oral examinations in the schools under their control. By 1845 the enrollments had become so large in Boston that the committee could no longer devote the time required for anything more than the most casual examination of each pupil with an oral quiz. To meet this situation the uniform written examination was adopted. The results were so gratifying that Horace Mann wrote the enthusiastic defense of written examinations to which reference has already been made.

In the latter part of the last century considerable pressure was being brought to bear from the outside upon the schools to make place for certain new and practical subjects such as manual training and home economics. But the school men opposed the move on the ground that there was hardly time to teach the subjects already in the curriculum. Then, in 1894, Dr. J. M. Rice had what he called an "inspiration." He says:[42]

> In truth, however, I came to recognize that this (the claims of school men following different courses of study) was all talk,—that no one really knew the facts, because there were no standards to serve as guides. Then one day, the idea flashed through my mind that the way to settle the question was to try it out. For a beginning I decided to take spelling, and on that very day I made up a list of 50 words with the view of giving them as a test to the pupils of the schools as I went on my tour from town to town. I have no record of the date of the inspiration, but I think it was some time in October, 1894

This was the origin of the important spelling inquiry which started a movement that not only has transformed the teaching of spelling but has brought to the fore a new technique for the settling of educational issues.

---

[41] Walter S. Monroe, "Educational Measurement in 1920 and in 1945," *Journal of Educational Research,* 38: 334–340, January, 1945.

[42] Quoted by Leonard P. Ayres, *op. cit.,* page 11. Quoted by permission of the Society.

The schools of every period have apparently had to meet the criticism that they are not so efficient as those "in the good old days." Usually, again, there is no defense except argument based upon mere opinions. The criticism was especially severe in the early years of the present century. Just at this time, in 1906, a fortunate event occurred which taught educators a second lesson in the value of comparative examinations. John L. Riley of Springfield, Massachusetts, discovered in an old attic a set of examinations which had been given in the Springfield schools in the year 1846. The thought occurred to him to give these same examinations to the pupils in the same city in 1906, just sixty years later.[43] In spite of the changes in the content of the subjects, he found the results distinctly favorable to the later schools. In ninth-grade spelling, for example, the pupils in 1846 had averaged 40.6 per cent, while the average was 51.2 per cent in 1906. In like manner, the geography average had risen from 40.3 per cent in 1846 to 53.4 per cent in 1906. But the greatest superiority was in the case of arithmetic, where the increase was from an average of 29.4 per cent in 1846 to 65.2 per cent in 1906. It was evident, therefore, that the facts were the most effective tools with which to meet criticism, and that comparative examinations were very useful in supplying these pertinent facts.

D. The History of Character and Personality Measurement

Crude beginnings. It is probably true that human beings began to pass judgment upon each other and to attempt evaluation of each other's character and personality long before the dawn of recorded history. But from the standpoint of measurement, these early efforts were both unsystematic and untrustworthy. Even when somewhat later these methods were reduced to systems, the results were still little better than chance. Examples of such prescientific systems which have exerted wide influence upon a credulous public are astrology, graphology, palmistry, and phrenology.

In spite of the antiquity of these attempts at evaluating personality and character, the scientific study of this field is comparatively new. Nor can it be said that the earlier pseudo-scientific approaches have ceased to influence the popular mind. Galton pioneered here as in so many other aspects of measurement. More than 60 years ago he came to the conclusion that "the character which shapes our conduct is a definite and durable 'something,' and that it is therefore reasonable to attempt to measure it."[44] He proposed rating scales with statistical analyses of results, and what he termed "rude experiments" suggested many later investigations. Without doubt Galton's ingenious suggestions marked the beginning of the scientific measurement of character.

---

[43] John L. Riley, *The Springfield Tests.* Springfield, Massachusetts: The Holden Patent Book Cover Co., 1908.

Later development. In recent years the analysis and measurement of personality and character have been greatly stimulated by the interest in educational and vocational guidance, mental hygiene, and character education.[45] It was soon recognized that success along these lines was conditioned upon the ability to measure other things besides general intelligence and academic achievement. With respect to character education, for example, one of the leaders in that field, Lentz, says: "Character education without character measurement would appear to be as logical as target practice in the dark, good shots and poor ones being equally gratifying."[46]

The first attempt to measure character by a test was probably that of Fernald in 1912, but the author's claims for the test were very modest.[47] Voelker, in 1921, devised some actual test situations for measuring character. By far the most ambitious attempt so far made is that of the Character Education Inquiry,[48] under the direction of Hartshorne and May, which extended over the five-year period, 1924–1929. These workers subjected all the promising tools then in existence to rigid trial, and devised many new and ingenious techniques of their own. Their main effort was directed at selecting representative and varied life situations which would afford a valid index of the totality of the character of the individual.

Most of these methods, however, had had interesting historical antecedents. For example, the celebrated physician, Galen, who lived in the second century, employed methods which were not unlike those in current use. On one occasion Galen, employed by the emperor to find out whether the empress was in love with a certain courtier, attempted to do so by having the suspected lover appear in the presence of the empress while the physician felt her pulse to determine the change in heart beat! It will be noted that Galen's technique suggests modern physiological methods of blood pressure, psycho-galvanometer, and the like, as well as the "sampling" method of the performance tests.

---

[44] Francis Galton, "Measurement of Character," *Fortnightly Review,* 42: 179–185, 1884.

[45] Some idea can be had of the extensive literature on the subject from examining the annual volumes of the *Review of Educational Research.* A selected bibliography on character tests, including 282 titles, selected from a complete list of about 1,000 titles, appeared in 1932. A supplementary bibliography for the next three years which appeared in 1935 included over 400 titles.

[46] Theodore F. Lentz, Jr., *An Experimental Method for the Discovery and Development of Tests of Character,* page 2. New York: Bureau of Publications, Teachers College, Columbia University, 1925.

[47] G. G. Fernald, "The Defective Delinquent Class Differentiating Tests," *American Journal of Insanity,* 68: 524–594, 1912.

[48] A complete report has been published by The Macmillan Company in three volumes.

It is doubtful whether, as a rule, tests of actual performance have sufficiently demonstrated their superior validity and reliability as measures of character and personality to justify the additional expense and inconvenience involved in their administration, except for purposes of research. In a recent summary Goodwin Watson, a recognized authority in personality evaluation, says: "It is probable that the last five years have brought some swing of psychological interest away from personality-test techniques and toward more emphasis upon the study of personality through ratings, anecdotal records, observation of behavior, and case studies."[49] In the final chapter of Symonds' comprehensive and analytical book, *Diagnosing Personality and Conduct,* the author makes this statement: "Probably the greatest usefulness will be found in ratings, the questionnaire, and the interview for obtaining evidence as to adjustments toward the environment, personal evaluation, attitudes toward reality, sexual relationships, morals, and feelings."[50] Recently the same author[51] notes clear evidence of a lag both in clinical research and in the translation of this research into educational practice. He says:

> Work which had been done in the decade from 1910 to 1920 on mental and achievement tests was being assimilated in educational practice in the 1920's. Basic work on the measurement and evaluation of personality that had been carried out in the 1920's was being assimilated in educational practice in the 1930's. Basic work on child guidance procedures, discussions of the meaning of mental hygiene in education, and the problems of pupil adjustment which were being investigated and elaborated in the 1930's is only now being assimilated in the practice of education in the schools of the nation.[52]

The historical development of rating scales, questionnaires, and interviews will now receive brief consideration. Strictly speaking, rating scales and questionnaires are only devices for recording the judgments of observers, rather than true measuring instruments.[53]

---

[49] *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II,* page 369. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1938.

[50] Percival M. Symonds, *Diagnosing Personality and Conduct,* page 567. New York: D. Appleton Company, Inc., 1931. Used by permission of D. Appleton-Century Company.

[51] Percival M. Symonds, "The Lag in Clinical Research," *Journal of Educational Research,* 38: 371–374, January, 1945.

[52] *Ibid.,* page 373.

Rating scales. The first rating scale in a modern sense was probably that of Galton for mental imagery, which was published in 1883. About the time of the appearance of the first Binet test, Karl Pearson proposed a scale for judging intelligence. One of the most famous scales specifically for measuring traits of personality is the Scott Man-to-Man Scale, introduced and extensively used during World War I. This scale has not been used widely since, however, partly because it is cumbersome to use, and partly because its validity has been seriously questioned by such leaders as Rugg and Thorndike. It remained for Hartshorne and May to restore somewhat the lost prestige of the rating procedure, partly by changing the name to "reputation measures," but mainly by improving the technique.[54]

The questionnaire. The invention of the much-maligned questionnaire is often ascribed to that versatile Englishman, Sir Francis Galton. But that the instrument was in existence at an earlier period in England, in fact if not in name, is evident from the following critical statement: [55]

> It is impossible to expect accuracy in returns obtained by circulars, various constructions being put upon the same question by different individuals who consequently classify their replies upon various principles.

But Galton undoubtedly improved and used extensively the questionnaire which was imported to America about 1880 by G. Stanley Hall. The instrument exists in many forms today, but its principal use in education is for measuring adjustment, attitude, and interest.

The Woodworth Personal Data Sheet began in 1917 as a method of measuring the ability of soldiers to adjust themselves to the trying conditions of army life. In 1923 Mathews adapted it for school use. The same year Cady published another revision which has been widely used with children in their teens. Two years later Laird made an adaptation for college use and provided a graphic rating scale for scoring. In 1919 Pressey published his widely used X-O Test, which is a sort of questionnaire covering a miscellaneous assortment of items having to do with emotionality. The questionnaire has also been used to measure other types of adjustment, such as introversion-extroversion by Marston and others, and ascendance-submission by Allport.

---

[53] B. Othanel Smith, *Logical Aspects of Educational Measurement,* Chapter I. New York: Columbia University Press, 1938.

[54] A historical note by the authors offers another illustration of the kinship between necessity and invention: "For a while it seemed that . . . rating scales as scientific instruments would be completely discarded. It was necessity that saved the day. While everyone talked about the superiority of objective tests, yet it was soon found that many qualities of character yield only stubbornly and expensively to objective testing. If character and personality studies were to continue, ratings had to be revived. In spite of all their difficulties, snares, delusions, and pitfalls, they are now staging a considerable 'comeback.'" *The Journal of Social Psychology,* 1: 66, February, 1930.

[55] Quoted from *Journal of the Statistical Society of London,* October, 1839, by Walter S. Monroe and Max D. Engelhart, *The Scientific Study of Educational Problems,* page 40. New York: The Macmillan Company, 1936.

The development of wholesome attitudes has been recognized in recent years as an important objective of education. Since about 1920 much attention has been devoted to the measurement of attitudes of various kinds. Hart's test of social attitudes and interests which appeared in 1923 and Watson's measurement of fairmindedness which appeared in 1925 are good examples. Beginning in 1928 Thurstone has been responsible for important improvements in the units of measurement employed in attitude questionnaires on many subjects. By having his questions scaled by a group of judges, Thurstone has found it possible to secure satisfactory results with fewer items. Figure 3 shows an adaptation of this weighting technique applied to a scale for measuring a pupil's attitude toward high school.[56] In this case the scale values range from .6 for Item 1 to 10.1 for Item 22. The pupil's score is the median scale value of the items checked.

The close relationship of interest to guidance, whether educational, vocational, or personal, has stimulated considerable activity directed toward its measurement. In 1907 G. Stanley Hall published a questionnaire study of the recreational interests of children which exerted a wide influence. This pioneer study has been followed by many others, of which the most extensive is probably that of Lehman and Witty, published in 1927. Moore appears to have been the first to use this technique for the measurement of vocational interests in 1921. Two other studies, employing somewhat different techniques, appeared in 1926. One of these, by Miner, offered paired comparisons, and the other, by Cowdry, employed a complicated scheme for weighting the scores. The best known of all is the Strong Vocational Interest Blank, which appeared in 1927.

The interview. One of the oldest forms of obtaining knowledge is the personal interview. It has always been an important tool in the hands of certain professional men, such as lawyers, doctors, and newspaper reporters; but it is used by everybody to some extent. Its chief value in education is probably in diagnosis and guidance.[57]

---

[56] For a discussion of this scale, see H. H. Remmers, G. C. Brandenburg, and F. H. Gillespie, "Measuring Attitude Toward the High School," *Journal of Experimental Education,* 2: 60–64, September, 1933.

[57] Thorndike has described the Stanford-Binet as "an approved, systematized and standardized interview." *The Measurement of Intelligence,* page 1. New York: Bureau of Publications, Teachers College, Columbia University, 1927.

The interview is used to supplement the ordinary objective evidence about a pupil, such as is afforded by his school record, by a firsthand knowledge of such things as his feelings and point of view. The

HIGH SCHOOL ATTITUDE SCALE[a]

Form B

Below is a list of twenty-five statements about school. Place a check mark before each statement with which you agree, and leave unmarked those with which you disagree. This test will in no way affect your standing in school.

_________1. School is like a prison.
_________2. I have a lot of fun in school.
_________3. School is all right.
_________4. My teachers always treat me fairly.
_________5. I like to go to school to be with other people.
_________6. Many of our great men have no high school education.
_________7. I hate most school work.
_________8. Some things about high school are all right.
_________9. High school training develops personality.
________10. High school is a good thing for some people and a bad thing for others.
________11. I would just as soon stay at home as go to school.
________12. All the better class of people have high school educations.
________13. The high schools lift the plane of sportsmanship in a community.
________14. Too much money is being spent on high schools for the benefit received.
________15. The high school teaches mostly old useless information.
________16. They won't teach things one really wants to know in high school.
________17. If one has plenty of money it may be all right to go to high school.
________18. I haven't any definite like or dislike for high school.
________19. Any old fogy knows more than a high school graduate.
________20. The kindest and best people I know don't have a high school education.
________21. High school cramps and dwarfs one's personality.
________22. America could not stand as a nation if it were not for our high schools.
________23. Our high schools teach immorality and indecency.
________24. High school training develops high ideals in pupils.
________25. High schools develop loyality.

[a] Prepared by F. H. Gillespie and published by Purdue University.

Figure 3. A Scale for Measuring Pupils' Attitudes Toward High School.

evidence indicates that the personal qualities of the interviewer are fully as important as the technique employed.[58]

Recent developments. Three promising recent developments relating to measurement require brief mention. The first of these is the attempt to subject personality to elaborate statistical analysis. Outstanding leaders in this movement have been Spearman of England, Thurstone of Chicago, and Kelley of Harvard. The second development is that of controlled observation or time-sampling. This is mainly employed in the child-study laboratories, notably those of Yale, Columbia, Iowa, and Minnesota. A third development has been in the direction of measuring public opinion; in this connection the success of the Gallup Poll has attracted wide attention. Recent volumes have described this third development.[59]

E. Some Important Publications

From the beginning, professional journals, books, and other publications have exerted a profound influence upon experimental psychology and the testing movement. Only the most important can be mentioned here.

Professional journals. The value of professional journals is that they keep the workers in one area continuously informed of what is going on elsewhere. The first psychological journal was *Mind,* founded in England by Bain in 1876. For eleven years it remained the only psychological journal in the English language, and so was the vehicle for most of the important psychological articles both in England and in America. One of the most important articles on measurement during the early years was probably that by Cattell entitled "Mental Tests and Measurements," which appeared in 1890 and contained some significant comments by Galton. The first psychological journal in America was *The American Journal of Psychology,* started by G. Stanley Hall in 1887. It has published many significant articles on measurement and statistics, but doubtless none more important than Spearman's "General Intelligence Objectively Determined and Measured," which appeared in 1904. This was the original formulation of the now well known two-factor theory of intelligence, and the beginning of "correlational psychology," which was influential in directing the attention of psychologists from faculties to factors.

---

[58] Valuable suggestions on the interview are found in: Percival M. Symonds, *op. cit.,* pages 450–484; and W. V. Bingham and B. V. Moore, *How to Interview* (Revised Edition), 308 pages. New York: Harper & Brothers, 1934.

[59] Hadley Cantril and Associates, *Gauging Public Opinion.* Princeton: Princeton University Press, 1944. 318 pages.

George H. Gallup, *A Guide to Public Opinion Polls.* Princeton: Princeton University Press, 1944, 104 pages.

For a critical summary and extensive bibliography see: Quinn McNemar, "Opinion-Attitude Methodology," *Psychological Bulletin,* 43: 289–374, July, 1946.

Hall also founded *Pedagogical Seminary,* which may be regarded as the first journal of educational psychology, in 1891. Three years later Binet started *L'Année Psychologique,* which was to be the principal agency for bringing his extensive work on the measurement of intelligence to the attention of the world. The *Teachers College Record,* started in 1900, has published many of Thorndike's important studies, and those of his students and coworkers. The first volume of *The Journal of Educational Psychology,* founded in 1910, contained an article by Huey on "The Binet Scale for Measuring Intelligence and Retardation," which was the first translation of the 1908 scale to appear in America. It has continued to be one of the most important journals on measurement. Two other journals, *School and Society,* and *Educational Administration and Supervision,* both founded in 1915, have included many reports on the use of tests both for research purposes and for the actual work of instruction and school administration. But no journal has been more important than the *Journal of Educational Research,* started in 1920. From the very first issue, which contained McCall's article on "A New Kind of School Examination," to the present time, it has exerted a wide influence upon the measurement movement.

Some important books. Some of the important books are briefly mentioned in chronological order, beginning with the pioneer period, which included roughly the first two decades of the century. In 1904 appeared E. L. Thorndike's *An Introduction to the Theory of Mental and Social Measurements,*[60] which made available for the first time to American students the statistical techniques necessary for educational research and measurement. Ten years later Truman Kelley's *Educational Guidance*[61] introduced educational workers to the alluring possibilities of partial and multiple correlations. Two important books appeared in 1916. One of these, Daniel Starch's *Educational Measurement,*[62] was the first book on achievement tests, and the other, L. M. Terman's *The Measurement of Intelligence,*[63] was the first adequate treatment of intelligence tests in the English language. In 1918 appeared the *Seventeenth Yearbook of the National Society for the Study of Education,*[64] Part II of which treats in some detail the history of the pioneer period in testing, and which gives descriptions of existing tests with suggestions as to their use. But it is probably most famous for containing Thorndike's statement, "Whatever exists at all exists in some amount," which has been accepted as a sort of creed by many workers in the field. The following year, in 1919, appeared Carl Seashore's *The Psychology of Musical Talent,*[65] a pioneer study in the measurement of specific intelligence, or aptitude in a restricted field.

---

[60] Published by Bureau of Publications, Teachers College, Columbia University.
[61] Published by Bureau of Publications, Teachers College, Columbia University.
[62] Published by The Macmillan Company, New York.
[63] Published by Houghton Mifflin Company, Boston.
[64] Published by Public School Publishing Company, Bloomington, Illinois.

Since the beginning of the third decade of the century, the "quantity production" stage referred to by Monroe has been achieved not only in the publication of tests but in books as well. Only a few of these can be mentioned here as representative of types. In 1922 appeared W. A. McCall's *How to Measure in Education,*[66] a comprehensive and critical book on achievement tests. The next year saw the publication of Ben D. Wood's *Measurement in Higher Education,*[67] the first treatise on measurement at the college level. In 1924 appeared G. M. Ruch's *The Improvement of the Written Examination,*[68] which was the first book wholly devoted to the new-type test. The year 1927 was especially productive, for at least five important books on measurement bore that date of publication. There were two notable books on intelligence, E. L. Thorndike's *The Measurement of Intelligence,*[69] and C. E. Spearman's *The Abilities of Man,*[70] each representing a distinct point of view. In 1927 also appeared the first two books specifically devoted to measurement in the high school, P. M. Symonds' *Measurement in Secondary Education,*[71] and G. M. Ruch's and G. D. Stoddard's *Tests and Measurements in High School Instruction.*[72] The same year Truman Kelley's *Interpretation of Educational Measurements,*[73] the most critical discussion now available of certain aspects of measurement, was published. The next year, in 1928, appeared another critical volume, Clark Hull's *Aptitude Testing,*[74] perhaps still the most complete discussion of that subject and destined to become one of the classics in the field of measurement.

Since 1930 there have appeared numerous books and monographs on the various phases of measurement and their application to the different educational levels. Many of these will be referred to in later chapters. It is too early, however, to attempt a just appraisal of their respective merits. Some evidence that measurement is coming of age is afforded by the fact that extensive bibliographies of tests and scales have appeared during this decade. The first edition of Gertrude Hildreth's *Bibliography of Mental Tests and Rating Scales*[75] was published in 1933. Three years later came Oscar Buros' *Educational, Psychological, and Personality Tests of 1933, 1934, and 1935,*[76] the forerunner of *The Mental Measurements Yearbooks,*[77] the first volume of which appeared in 1938. The publication of this critical volume marked an important milestone in the history of educational measurement.

---

[65] Published by Silver, Burdett and Company, New York.

[66] Published by The Macmillan Company, New York.

[67] Published by World Book Company, Yonkers.

[68] Published by Scott, Foresman & Company, Chicago.

[69] Published by Bureau of Publications, Teachers College, Columbia University.

[70] Published by The Macmillan Company, New York.

[71] Published by The Macmillan Company, New York.

[72] Published by World Book Company, Yonkers.

[73] Published by World Book Company, Yonkers.

[74] Published by World Book Company, Yonkers.

F. Some Recent Tendencies

Test construction. The "quantity production" stage of test construction in America now seems definitely past. The emphasis has turned to quality rather than mere quantity, although it is still too much to say that a recent copyright date on a test is ample assurance of high merit. Kelley's observation made in 1927 that "the ruts of the test movement are already so deep that there are many who do not see beyond them"[78] is still, unfortunately, true. However, test makers as a group no longer unblushingly make the enthusiastic claims for their products that were common even a decade ago. Instead there has grown up a more critical and becomingly modest attitude, which is probably the most characteristic feature of the present trend. One alert observer[79] as early as 1920 noted "evidences of the beginnings of a critical attitude toward educational tests."

Another tendency, largely an outgrowth of this critical attitude, is to extend the field of measurement into new areas and to develop new, and usually more *specific,* types of tests. For example, instead of a primary interest in developing tests of general intelligence, the emphasis is upon developing tests of specific intelligence along particular lines. Reading readiness and other aptitude tests are representative of the trend. Even the so-called general intelligence tests that have appeared during the past twelve or fifteen years

---

[75] Published by The Psychological Corporation, New York

[76] Published by Rutgers University Press.

[77] Published by Rutgers University Press.

[78] Truman Lee Kelley, *Interpretation of Educational Measurements,* page 16. Yonkers: World Book Company, 1927.

[79] Walter S. Monroe, "Educational Measurement in 1920 and in 1945," *Journal of Educational Research,* 38: 334–340, January, 1945.

"attempt to be more diagnostic of special abilities and disabilities." [80] Increased attention to the reliability, and more particularly to the validity, of tests and individual test items is also a notable trend. The result has been the appearance of new types of test forms and test situations. Along with this has come the realization that standard tests do not fully meet all the needs of measurement, and that in consequence greater emphasis must be placed upon the development of improved techniques for constructing informal teacher-made tests and other techniques of evaluation.

Monroe makes the following excellent summary [81] of the situation:

> The most significant trends appear to be (1) the attack upon the consequent discrediting of essay examinations, (2) the development of objective tests, and the emphasis upon reliability as the criterion by which measuring instruments were evaluated, and (3) the development of diagnostic and prognostic uses. What of the future? Any attempt to project lines of development into the future is attended with uncertainty. But, if I interpret correctly current educational writings in this field, three trends are indicated: (1) a growing emphasis upon validity and a consequent decreasing emphasis upon reliability as the criterion for evaluating measuring instruments; (2) a decline of the faith in indirect measurement, and an increasing emphasis upon direct measurement as a means of attaining satisfactory validity; and (3) a growing respect for essay examinations as instruments for measuring certain outcomes of instruction.

Use of tests. A British psychologist [82] has suggested that a new scientific technique seems to go through three stages, as follows:

> The first is the early stage of development when no one, except its inventors, is interested in it, and those working on other lines regard it with indifference or suspicion or else think it silly. In the second stage it begins to gain support, and in the third stage everyone wants to use it whether they understand it or not. There is then danger of a fourth stage of disillusionment, and this is the time for critical examination.

In the case of standard tests in America, the stage of indifference and suspicion, with which Rice's spelling inquiry was met, had largely passed when the first standardized tests appeared during the first decade of the present century. Since that time there have been three rather clearly defined stages, which may be designated as those of curiosity, confidence, and critical caution.

---

[80] Harry J. Baker, "Intelligence and Its Measurement," *Review of Educational Research,* 5: 198, June, 1935.

[81] Walter S. Monroe, "Some Trends in Educational Measurement," *Twenty-Fourth Annual Conference on Educational Measurements,* page 35. Bulletin of the School of Education, Indiana University, Vol. XIII, No. 4. Bloomington, Indiana: Bureau of Coöperative Research, 1937.

[82] J. O. Irwin, "Correlation Methods in Psychology," *British Journal of Psychology,* 25: 86–91, July, 1934.

The first stage was that of *curiosity.* In this stage teachers and school officials tried out tests merely because they were something new and because their use gave evidence, if indeed superficial in character, of up-to-date-ness. This attitude tended to die a natural death as the novelty wore off.

The second stage was that of *confidence,* or in some instances that of overconfidence. Standard tests were "swallowed, hook, line, and sinker." Test results were uncritically accepted at their face value. IQ's were naïvely taken as accurate measures of innate capacity wholly apart from environmental opportunities, and so were as fixed as the laws of the Medes and the Persians.[83] In like manner, achievement test scores were accepted as fully adequate measures of the important outcomes of instruction. If only such tests were objective, they were assumed to be sufficiently accurate for valid comparisons, not only of one school or class with another, but also of one pupil with another, or even of one aspect of a pupil's achievement with another aspect of his achievement.[84] There is some evidence that this attitude is on the decline, although unfortunately it is still found too often in certain quarters.

The third stage may be termed that of *critical caution.* While by no means universal, this more wholesome attitude, on the whole, characterizes the present. Hildreth points out some beneficial results of this change: "A more critical attitude toward intelligence measurement, as the outcome of continued experimentation, has resulted in more authoritative research findings, more sensible and intelligent interpretation of data." [85] This attitude has shown itself with respect to achievement tests and personality measurements as well. The result has been not so much the curtailment of the use of tests, as their more critical use and the more cautious interpretation of test scores. On the whole, the emphasis at the present time is on the use of standard tests, not so much for comparative purposes, as to provide a basis for guidance and remedial instruction. It is increasingly recognized that tests are means and not ends, and that even the best test is but a tool, the value of which depends upon the skill and the intelligence with which it is used.

---

[83] What happened to intelligence testing following World War I has been described as follows: "As many of the subjects tested were children of school age, because Binet's scores gave a good correlation with ability for school work, and perhaps because of the relative simplicity and economy of the methods, mental testing was oversold, and careful psychological work in the field of individual differences still suffers from this effect." Francis N. Maxfield, "Trends in Testing Intelligence," *Eduational Research Bulletin,* 15: 137, May 13, 1936.

[84] What happened to achievement testing has been described as follows: "In the widespread use of objective tests at the high school and college levels, there is apparent a child-like faith in the efficacy of objective tests as instruments for measuring school achievement. . . . A little knowledge has become a dangerous thing." Walter S. Monroe, "Hazards in the Measurement of Achievement," *School and Society,* 41: 48–49, January 12, 1935.

[85] Gertrude Hildreth, "Applications of Intelligence Testing," *Review of Educational Research,* 5: 200, June, 1935.

This enlarged and more critical attitude on the part of enlightened school officials has been well stated by Maxfield:[86]

> In problems of school administration the massed data from intelligence tests will be interpreted by statistical methods. In dealing with the problems of individual pupils the case-study method of the clinical psychologist will prevail. Inventories of personality, scales of social adjustment, and the like, will supplement tests of intelligence. Diagnostic tests will be supplemented by diagnostic teaching. But no synthesis or interpretation will be attempted without knowledge of the pupil's physical condition, his home background, his previous school history, his vocational interests, his social and emotional reactions, and the like. The weight given in this synthesis to scores on intelligence tests will vary with the problem presented. The case-study method can be adapted to any philosophy of education and to any educational aims and objectives.

SELECTED REFERENCES FOR FURTHER READING

Ayres, Leonard P., "History and Present Status of Educational Measurements," *Seventeenth Yearbook of the National Society for the Study of Education, Part II.* Bloomington, Illinois: Public School Publishing Company, 1918. Chapter I.

Boring, Edwin G., *A History of Experimental Psychology.* New York: D. Appleton-Century Company, 1929. 699 pages.

Boynton, Paul L., *Intelligence: Its Manifestions and Measurement.* New York: D. Appleton-Century Company, Inc., 1933. Chapters V and VI.

Freeman, Frank N., *Mental Tests. Their History, Principles and Applications* (Revised Edition). Boston: Houghton Mifflin Company, 1939. Chapters I–VIII.

Greene, Edward B., *Measurements of Human Behavior.* New York: The Odyssey Press, 1941. Part II.

Odell, C. W., *Educational Measurement in High School.* New York: D. Appleton-Century Company, 1930. Chapter II.

Peterson, Joseph, *Early Conceptions and Tests of Intelligence.* Yonkers: World Book Company, 1925. 320 pages.

Pintner, Rudolph, *Intelligence Testing, Methods and Results* (New Edition). New York: Henry Holt & Company, 1931. Part I.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman & Company, 1929. Chapters I and III.

Smith, B. Othanel, *Logical Aspects of Educational Measurement.* New York: Columbia University Press, 1938. Chapters I, II, and III.

Stoddard, George D., *The Meaning of Intelligence.* New York: The Macmillan Company, 1943. Part II.

Young, Kimball, "The History of Mental Testing," *Pedagogical Seminary,* 31: 1–48, March, 1924.

---

[86] Francis N. Maxfield, "Trends in Testing Intelligence," *Educational Research Bulletin,* 15: 140, May 13, 1936.

CHAPTER III

The Characteristics of a Satisfactory Measuring Instrument

A. Introduction

Importance of the problem. What are the earmarks of a good test, examination, or other measuring instrument? In the selection of a test, as in the selection of an automobile, it is important to know what to look for. There is usually a choice among many possibilities which are very unequal in merit. Each year many automobiles are bought because of the appeal of some gadget, such as a fancy radiator ornament or cigarette lighter; and many standard tests are bought for no better reason. Whether a purchaser *buys,* or is merely *sold,* depends largely on whether or not he knows what to look for in the article in question. Moreover, every teacher will have occasion to use tests of his own construction, and should know what qualities to strive for in such tests. As a rule, the same characteristics are essential in an informal test made by the classroom teacher as in a standard test bought ready-made from a publisher.

In any satisfactory measuring instrument three qualities are indispensable. These are:

1. Validity
2. Reliability
3. Usability

It is essential, therefore, that every teacher have a clear idea regarding the meaning of these characteristics, and know how to judge their presence in tests, whether standardized or nonstandardized.

B. Validity

Meaning of validity. By validity is meant the degree to which the test or other measuring instrument measures what it claims to. In a word, *validity* means *truthfulness.* Does the test really measure what it purports to? For example, whether a so-called "arithmetic reasoning test" is valid or not depends upon the extent to which it succeeds in measuring reasoning ability in arithmetic rather than other things, such as reading ability or general intelligence. Validity, then, refers to the truthfulness of the test and is always its most important characteristic. No matter what other merits the test may possess, if it lacks validity, it is worthless. Whether you are selecting a standard test or making an informal test, the first thing to consider is its validity. How, then, does one judge whether or not a test or other measuring instrument is valid?

General considerations. The answer to this question may best be approached by giving attention to some preliminary considerations of a general nature.

1. The *nature of the thing being measured* must always determine the methods and materials of measurement. In order to judge the validity of an intelligence test, for example, it is necessary to consider what intelligence is, what its qualities are, or at any rate, how it manifests itself. In like manner, in order to judge the validity of an achievement test, it is necessary to consider what it is that the achievement test is supposed to measure. This means that the first step in judging the validity of achievement tests is some authoritative statement of the specific objectives of the course or subject.
2. Any measurement in education is always a *sampling,* never entirely complete. The test maker relies upon a sample much as does the chemist in the health department in passing upon the quality of the city's water supply. In psychological language, any test is merely a series of situations designed to call forth a sufficient number of representative responses to enable the examiner to determine the amount of the thing in question that happens to be present.
3. The *accuracy* of the measurement, its fineness of discrimination, will depend upon the purpose it is to serve. A cheap alarm clock will usually suffice for a housewife in determining when to prepare lunch or to expect the postman, but a finer timepiece is required for the locomotive engineer. In like manner, a sundial or hour glass may be adequate for a gardener, but a split-second watch is essential for a football official. It would be almost as absurd to attempt to use a sundial to time a football game as to use it for measuring temperature or wind velocity. In other words, the validity of the measuring instrument must always be considered in relation to the purpose it is to serve. Validity is always specific, in relation to some definite situation. A test is not just valid; it is *valid for something.* There is no such thing as general validity.

I. The Validation of Intelligence Tests

Although the job of constructing the so-called tests of general intelligence is usually turned over to the specialist, a general knowledge of how such tests are validated will enable the teacher to select and use them more discriminatingly.

The meaning of intelligence. What, then, is meant by "general intelligence," the thing such tests claim to measure? Although there is no unanimity among psychologists regarding the exact definition of intelligence, there is substantial agreement that what existing tests attempt to measure is the innate capacity to learn, particularly to learn the academic tasks imposed by the school. Such a conception of intelligence is not very "general," after all. It is clearly narrower than the popular notion, since it is restricted largely to abstract intelligence and leaves out of account social intelligence, mechanical intelligence, and intelligence in special fields such as athletics, music, or oratory.

It is also clear at the outset that innate intelligence can be measured only indirectly; its presence must be inferred from the observed behavior of the individual, his reactions to certain carefully chosen and controlled situations called tests. Such tests should meet two general requirements: First, there must be a sufficiently large and varied assortment of test situations to call forth a wide variety of mental operations, primarily of the higher type, such as imagination, judgment, and reasoning. Second, the situations must be of such a nature that every individual taking the test has had approximately equal opportunity to learn, and as far as possible, equal motivation. This second standard is hard to meet and is usually only approximated even in the best tests. It clearly rules out tests that involve special talents such as for music or art, and makes questionable those that depend on specific school experience, which is by no means uniform for all pupils. In general, group tests meet these standards, especially the second, less well than do individual tests. The Army Alpha, for example, not only employs such situations as reading vocabulary and arithmetic, but, being originally designed for soldiers, has material that is more within the experience of men and boys than of women and girls.

The Terman criteria. In developing the well known Stanford Revision of the original Binet Scale, Terman relied upon three additional criteria of intelligence: namely, age increase, coherency, and world success.[1] *Age increase* means that each test item must show an increasing percentage of successful responses from one year level to the next. This is only a partial criterion, since it must assume that the items chosen are of a type that may reasonably be expected to measure intelligence. Purely physical measurements, for example, such as strength of grip, or speed in running, show age increases. The second criterion, *coherency,* is based on the assumption that the whole test is a more valid measure of intelligence than is any of its parts. Upon the basis of the entire test, the group is divided into dull, normal, and bright sections of approximately equal size. Then, to be acceptable, each item must discriminate within the sections by showing a progressively increasing percentage of successes as we go from dull to bright. This procedure really measures the internal consistency of the test, much as a logician judges the validity of a course of reasoning. Both Galton and Binet used the method of contrasting groups, although their groups were selected from external criteria rather than from the test itself. As an English author observes: "Thus in philosophy, as in science, the only test of validity is self-consistency." [2]

---

[1] For a discussion of the procedure used in the Revised Stanford, see: Quinn McNemar, *The Revision of the Stanford-Binet Scale,"* pp. 1–14. Boston: Houghton Mifflin Company, 1942.

The third criterion, *world success,* is the ordinary common-sense standard of everyday life. As the test is validated on children, this really means the child's world, which is primarily that of the school, his standing in which is reflected in his academic record. This is, of course, not a perfect criterion. It is not only highly subjective in character, but it throws the primary responsibility ultimately upon the judgment of teachers, which, because of its limitations, the test is being designed to replace or supplement. This is not so bad as it seems, however, for the basis is not that of the pupil's mark on a single examination, whose notorious unreliability has already been described, but rather that of his *entire record* for an extensive period, a far more stable thing. Furthermore, the reliance is usually placed not upon the judgment of any single teacher, but rather upon the average of several experienced teachers. The consensus of competent persons is the ultimate criterion of values from the constitutionality of a law down to the beauty contest at the local theater.

Individual versus group tests. It is generally assumed that an individual test is likely to meet more fully the criteria described than does a group test. Furthermore, the individual test permits the trained examiner to observe more carefully the behavior of the subject during the course of the examination. For example, if the subject shows signs of nervousness or refuses to co-operate fully, the examiner realizes that a valid measure of intelligence is impossible under the circumstances, and so waits for a more opportune time. Also, if the subject is handicapped by defective vision or hearing, this condition is almost certain to be discovered by the examiner, who then takes it into account in making his interpretations and recommendations. For these reasons the individual intelligence test is usually taken as the criterion or standard for validating the group test. In most instances the Stanford Revision of the Binet Scale is the one used. Sometimes one group test is validated by comparing the scores made on that test with those made by the same individuals on another group test, or possibly with some combination of two or more group tests. For all such comparisons with a criterion, whether it be the Stanford Revision or with some group test or tests, the Pearson product-moment coefficient of correlation is usually employed. It is then referred to as the *validity coefficient.* If the agreement with the criterion is perfect, the coefficient is 1.00, and if there is no consistent relationship at all, the coefficient is .00. Naturally the nearer the coefficient approaches 1.00, the higher the validity is said to be, although in the last analysis everything depends upon the validity of the criterion itself.[3] Usually the most difficult step in test validation is the determination of an adequate criterion.

---

[2] William C. D. Dampier-Wheham, *A History of Science and Its Relations with Philosophy and Religion,* page 466. New York: The Macmillan Company, 1930.

Table 3 illustrates various types of correlation data used for studying the comparative validity of group intelligence tests to be used for a specific purpose. Price[4] undertook to determine which of four widely used tests was the most valid for classifying pupils at the entrance to the ninth grade, or the freshman year of the four-year high school in Kentucky. The data regarding the validity of the various tests have usually lumped rural and city schools together and in no case have been based on Kentucky

TABLE 3

CORRELATION OF FOUR GROUP INTELLIGENCE TESTS WITH THREE CRITERIA OF VALIDITY (AFTER PRICE)

VARIABLE	DETROIT ADVANCED	KUHLMANN-ANDERSON	TERMAN GROUP	MCCALL MULTI-MENTAL
Detroit Advanced	—	.88	.86	.81
Kuhlmann-Anderson	.88	—	.86	.77
Terman Group	.86	.86	—	.78
McCall Multi-Mental	.81	.77	.78	—
Average of other three tests	.91	.88	.90	.83
Average of freshman marks	.46	.33	.56	.42

---

[3] For a fuller discussion of correlation, see pages 237–246.

[4] Orville Kelly Price, *Comparative Validity and Reliability of Four Intelligence Tests in the Ninth Grade,* Master's Thesis, University of Kentucky, 1933.

children. The procedure adopted by Price was to give all four tests to the ninth-grade pupils in several Kentucky high schools at nearly the same time in the school year. It was then possible to determine the extent of agreement of each test with every other test and with a combination of the other three tests. To make the scores comparable, they were reduced to the IQ as a common unit. A second criterion was the correlation of each test with the average of the other three. As one purpose of intelligence tests in the high school is to predict future academic achievement, the average mark of the pupils at the end of the year was another criterion. It will be noted that the group tests, as a rule, agree more closely with each other in combination than individually, and that the lowest coefficients are with teachers' marks.

The point to be stressed here is that not all validation procedures give the same results, or necessarily rank the tests in the same order. One must consider, therefore, not only the size of the validity coefficient but the procedure used as well. The same thing is also true of aptitude tests, achievement tests, personality inventories, or other measuring instruments. O'Rourke, for example, says of aptitude tests: [5]

> Reports on the predictive value of the same aptitude test range from low to high, and give us little basis for judging the comparative value of tests or even the validity of a single test. An explanation of the criterion and the factors used in it should accompany every report of a test program.

II. The Validation of Achievement Tests

Curricular versus statistical validity. In some respects the validation of an achievement test is more difficult than the validation of an intelligence test, and a greater number of procedures are employed for its determination. In discussing the validation of achievement tests a distinction should be made between *curricular* validity and *statistical* validity. By curricular validity is meant the extent to which the content of the test is truly representative of the content of the course. Curricular validity implies an act of judgment as to the adequacy of the sampling included in the test. In the earlier days this was interpreted to mean merely the extent to which the items of the test included a representative sampling of the essential materials employed in instruction. More recently, however, curricular validity is thought of, not primarily in terms of subject matter, which at best is merely the stimulus, but rather in terms of the *mental reactions* expected of the pupils themselves.

---

[5] L. J. O'Rourke, "Vocational Aptitude Tests," *Review of Educational Research*, 8: 268, June, 1938.

In other words, the center of gravity has shifted from the curriculum to the child.

Statistical validity refers to the mathematical processes for determining the degree to which the test agrees with, or correlates with, some criterion which is set up as an acceptable measure of the thing in question. Some of these statistical procedures aim at validating the test as a whole and others at validating the items individually. Although the procedures commonly employed by professional test makers are often rather technical, especially for item validation, the essential ideas are relatively simple. The validation of the test as a whole will be considered first.

TABLE 4

DISTRIBUTION OF METHODS OF VALIDATING 183 STANDARD ACHIEVEMENT TESTS (AFTER PETERS AND CROSSLEY)

METHOD	NUMBER OF TESTS
Textbook analysis	62
Correlation with school marks	47
Pooled judgment of competent persons	46
Items selected for difficulty	42
Determination of social utility by word counts	34
Correlation with previously validated measures	28
Logical or psychological analysis	26
Determination of social utility by error counts	18
Increases in percentage of successes with successive ages or grades	15
Empirical tryout	15
Opinion of author	15
Analysis of courses of study	14
Determination of social utility by scientific job analysis	13
Correlation between parts of the test, each part testing for different features	9
Determination of social utility on basis of general aims	9
Determination of social utility by frequency of references in reading	6
Analysis of final examination questions	6
Differential scores of spaced groups	6
Correlation with teacher's ratings	4
Agreement with observation of individuals	4
Questions submitted by competent persons	3
Correlation with college entrance examination scores	1

Test validation methods. Peters and Crossley[6] made a study of the validation methods employed in constructing "all the more important achievement tests, so far as information was available," that had appeared prior to December, 1928. When the desired information was not available in print, it was sought directly from the author himself. Even this source failed to yield information for fifty of the tests, which is a fact of significance in itself, as many of the letters from the authors were extremely vague or entirely off the subject. The analysis of the methods employed in validating the 183 tests for which the information was available is given in Table 4. Several points are worth noting. The wide diversity of practice reported shows that at the time the study was made no one technique had been accepted in practice as distinctly better than any other. If, as Peters and Crossley suggest, it is reasonable to add to the 15 who frankly admitted that the criterion used was the author's judgment, the 50 whose answers were vague or beside the point, it is rather disturbing to find that the most common method, the one used in more than a fourth of these tests, was the personal opinion of a single individual. Of course, there will always be a place in test construction for human judgment, especially the pooled judgment of a group of competent persons.

---

[6] Charles C. Peters and Elizabeth Crossley, "The Relation of Standardized Tests to Educational Objectives," *Second Yearbook of the National Society for the Study of Educational Sociology,* pages 148–159. New York: Bureau of Publications, Teachers College, Columbia University, 1929.

The technique employed in the preparation of the Cooperative Achievement Tests represents an effective combination of statistical analysis and the judgment of experts. These tests are constructed by a trained staff working in close co-operation with classroom teachers, subject-matter specialists, and test technicians. The procedure is outlined as follows: [7]

a. Preliminary planning and selection of content.
    Analyses of curricula, textbooks, research studies, etc.
    Formulation of objectives and determination of general plan
    Preparation of detailed test outlines based upon survey of materials
    Submission of outlines to authorities for criticism
    Revision of test outlines in accordance with suggestion of critics
b. Preparation and editing of test items.
    Writing of items by test editors and cooperating experts
    Submission of items to authorities for criticism
    Revision of items in view of suggestions received
    Preparation of experimental forms of test
c. Administration of experimental forms to a representative sampling of students to obtain item difficulty and validity indices, and to detect items which may be weak or ambiguous.
d. Preparation of final form.
    Selection and revision of items for tentative final form
    Obtaining from experts in subject-matter fields, test technicians, etc., suggestions and criticisms of the tentative final form
    Revision and final editing of the test, based on the criticisms and suggestions received
e. Administration of final form of test with earlier forms for equating and determination of scaled scores.

---

[7] *Cooperative Achievement Tests for High School and College Classes,* page 5. New York: Cooperative Test Service, 1945.

Perhaps attention should also be called to certain limitations of frequency of mention or use as a criterion for selecting materials either for the curriculum or for the test. In the first place, to accept *what is* as a criterion of *what ought to be* leaves no room for progress. For example, someone has defined a synonym as a word you use when you do not know how to spell the word you want. It can scarcely be doubted that there is a wide margin between the words actually used in ordinary speaking and writing and those that should be used to convey best the meaning intended. In the second place, frequency by its very nature is a poor standard for judging importance. For example, birth and death occur but once in the life history of an individual, and yet who would say they are for this reason less important than dressing and undressing, which occur every day? Frequency of use, therefore, although doubtless important as one measure of social utility, can rarely be regarded as the best criterion for validating a test. It should usually be employed with other criteria, rather than alone.

Some criticisms of test validity. One of the commonest criticisms of the validity of achievement tests, especially those of the objective type, whether standardized or nonstandardized, is that they are predominantly factual in character. It is alleged that they succeed merely in measuring verbal memory as distinguished from genuine understanding, and leave unmeasured the really important outcomes such as discrimination, judgment, intellectual and emotional attitudes, appreciations, and the ability to make intelligent application of knowledge to new situations. Even the best friends of achievement tests will readily admit that *as such tests are commonly made and used, the criticism has merit.* In fact, no one has recognized the limitation of existing tests more clearly than some of the outstanding leaders of the measurement movement itself. A quarter of a century ago Thorndike wrote:[8]

> In the elementary schools we now have many inadequate and even fantastic procedures parading behind the banner of educational science. Alleged measurements are reported and used which measure the fact in question about as well as the noise of the thunder measures the voltage of the lightning. To nobody are such more detestable than to the scientific worker with educational measurements.

More than a decade ago Monroe[9] wrote about the "child-like faith in the efficacy of objective tests as instruments for measuring school achievement" on the high-school and college levels. Three examples of unwarranted beliefs were cited:

---

[8] E. L. Thorndike, "Measurement in Education," *Twenty-First Yearbook of the National Society for the Study of Education, Part I,* page 8. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1922.

[9] Walter S. Monroe, "Hazards in the Measurement of Achievement," *School and Society,* 41: 48–52, January 12, 1935.

> I. Objectivity in scoring is an essential requirement for a satisfactory test, and if a test is objective, the scores yielded by it may be considered highly accurate measures of school achievement.
>
> II. If a test has been shown to be highly reliable, the scores yielded by it are highly accurate measures of the achievement specified by its announced or implied function.
>
> III. A high correlation with a criterion is sufficient evidence to justify the use of the scores yielded by a test as highly accurate measures of the achievement considered to be defined by the criterion.

While all three points are related to validity, the first two are more appropriately treated in later sections. The third point merits further discussion here.

Validity coefficients are no exception to the general principle which holds that all coefficients of correlation are definitely influenced by the variability of the group. For any given type of data a heterogeneous group gives consistently higher coefficients than a homogeneous group, regardless of other considerations. This phenomenon alone makes it impossible to regard the validity coefficient as absolutely fixed in magnitude. For example, a coefficient of .60 for a single grade may actually be more significant than one of .90 for several grades thrown together. Monroe points out that such coefficients also vary with the "general plan of the curriculum and the objectives toward which the instruction is directed." And, of course, the value of the validity coefficient always depends ultimately upon the validity of the criterion itself. After a test has been once used, it tends to influence both the teaching of the instructor and the study of the students, so that "a given type of objective test is likely to become less and less valid as its use is continued." In view of these facts most studies comparing the validity of one test with that of another are inconclusive, especially so when based on different groups. Comparisons of the relative validity of essay and new-type tests and of various forms of new-type tests are still more risky, for it is usually impossible to tell whether all forms were equally appropriate for the objectives in question, were constructed with equal skill, or aroused equal degrees of motivation. Such considerations lend support to Monroe's conclusion that the "coefficient of validity calculated for a test is a statistic of uncertain value." However, one thing is certain: Validity coefficients can rarely be taken at face value.

Tyler's suggestions. The upshot of the above discussion is that in judging the validity of an achievement test, for the present at least, *major dependence must be placed, not on the statistical anal-*

ysis of test results, but on the logical and psychological analysis of test construction. On this point no criticism in recent years has been more influential than that of Ralph W. Tyler of the University of Chicago.

The Tyler technique of test construction is based upon a broad conception of validity. He regards a valid test as one which affords satisfactory evidence of the degree to which the students are actually reaching the desired objectives of teaching, these objectives being specifically stated in terms of the kinds of behavior expected in the students. Tyler summarizes the process as follows: [10]

> All methods of measuring human behavior involve four technical problems: (1) defining the behavior to be evaluated, (2) selecting the test situations, or determining the situations in which the behavior is expressed, (3) developing a record of the behavior that takes place in these situations, and (4) evaluating the behavior recorded. Regardless of the type of appraisal under consideration, whether it be the observation of children at play, the written examination, the techniques of the psychological laboratory, the questionnaire, or the personal interview, these problems are encountered. The choice of the methods of measurement rests primarily upon the effectiveness with which the methods solve these problems in the particular case under consideration.

A few examples may be helpful. In chemistry the objectives sought by the instructors would doubtless include such abilities as understanding technical terms, remembering important facts and principles, applying important chemical principles to concrete situations, expressing chemical relationships by appropriate equations, and acquiring certain laboratory skills. In like manner, the objectives in English might include the effective use of English in speaking and writing, acquaintance with certain literary masterpieces, critical skill in evaluating the major types of literary productions, and the appreciation of good literature. These illustrations are sufficient to make it clear that no one type of test situation can measure adequately such a variety of teaching objectives. In fact, the outstanding weakness of both standardized and nonstandardized tests is that they have assumed that the attainment of one objective, usually knowledge, is sufficient assurance of like accomplishments in the others. Tyler has shown that this assumption is quite as erroneous as that made by Cattell and Galton, who in the pre-Binet stage of mental measurement accepted sensory discrimination as indicative of the higher mental processes, such as judgment and reasoning. He found, for example, that in elementary biology the correlation between information and the ability to apply principles was .40, that the correlation between information and the ability to interpret experiments was .41, and that the correlation between information and skill with the microscope was only .02. Even when allowance is made for the homogeneity of the groups involved, the relationship is far from close. Until recently most tests have apparently been built on the assumption that such correlations are around 1.00.

---

[10] *Thirty-Fourth Yearbook of the National Society for the Study of Education,* page 114. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1935.

Direct versus indirect methods. It has been found over and over again that the way to attain an educational objective in teaching is to aim at it directly rather than to rely upon transfer of training to bring it about indirectly. Apparently the same thing is true of testing also; the way to measure the extent to which an educational objective has been realized is to aim at it directly wherever possible, rather than to infer its existence from the indirect measurement of something else.

But this principle has often been violated in educational practice. We have been content with indirect measurement when we might have had direct. The Ayres educational index, for example, was based, not directly upon the educational *achievement* of the several states, but indirectly upon the educational *opportunities* offered, as measured by money expended, length of school term, and the like. Even today the important state and regional accrediting associations of colleges and secondary schools do not as a rule aim directly at the *product* of these schools, in terms of desired changes in pupil behavior, but rather indirectly at the *facilities* offered, such as size of classes, degrees held by members of the faculty, and number of volumes in the library. In view of the low correlation between intelligence and school achievement, which averages below .50, and that between the possession of knowledge and the ability to use it, which averages even less, one would hardly expect to find the perfect correlation between educational facilities and educational performance that such practice appears to assume. It is doubtless still true that there are Mark Hopkinses capable of transforming mere logs into colleges, while marble palaces may remain but piles of stone for lack of such a magic touch. In any case, it is *safer* to examine what is happening to the student at his end of the log, than to remain content to measure the dimensions of the log, or even the credentials of the individual who happens to be at the other end.

There is considerable experimental evidence which reveals the inadequacies of such indirect measures. For example, Trimble and Remmers[11] found that 11 per cent of the pupils in the high schools rated by the Indiana accrediting association as second class, and 9 per cent of those in the high schools rated as third class made scores on a comprehensive test in mathematics that were as good as or even better than any scores made by pupils in the schools rated as first class. Jessup[12] refers to the state-wide survey of high-school achievement in Iowa, where there was found an overlapping of more than 90 per cent between the schools that were accredited by the North Central Association and those that were not. He also cites a study which found that only 15 per cent of the colleges whose students performed best in graduate schools were on the approved list of the Association of American Universities, whereas 23 per cent of those whose students ranked lowest were on the approved list.

---

[11] Otis C. Trimble and H. H. Remmers, *Measures of Educational Outcomes Versus Standards of Institutional Machinery as High School Accrediting Criteria*, 37 pages. Purdue University Studies in Higher Education. Vol. 22, March, 1933.

It may, of course, be expedient at times to rely upon indirect evidence, but at any rate, one should do so only where direct evidence is not available and even then with full realization of the risks involved. For example, up to the present time test makers have found it difficult to devise suitable instruments for measuring such intangible outcomes of teaching as attitudes, appreciations, and interests. But it is probably true that the better standard tests come far closer to measuring the objectives actually attained or aimed at in educational *practice* than they come to measuring those suggested as desirable in educational *theory.* It is apparently just as difficult to *teach* these intangible things as it is to *test* them. It seems reasonable to think that it is no less difficult to provide the appropriate teaching materials for bringing about the right kind of attitudes, appreciations, and interests than it is to provide the appropriate testing materials for determining how well the job is being done. It should be kept in mind that a valid test consists largely of a *representative sampling of the materials that make up the course.* It should help to clarify the atmosphere once and for all to recognize frankly that *the less tangible outcomes are harder to teach and to test than the more tangible outcomes.* And it may be that for some time to come we shall have to be content to aim at both indirectly.

But this is no permanent solution to the problem. One of the important services the measurement movement can render education is the clarification of its objectives. The necessity for this should be apparent both to the curriculum maker and to the test maker. Considerable progress has already been made in this direction, and more will doubtless be forthcoming. The work of Wrightstone[13] is an illustration. He reports a series of tests in the social studies with such a diversity of aims as the interpretation of facts, the making of generalizations, the organization of data, several important work-study skills, and certain civic attitudes and beliefs. Recent reports of the Eight-Year Study of the Progressive Education Association indicate substantial progress in this direction.[14]

---

[12] Walter A. Jessup, "The Integrity of the American College from the Standpoint of Administration," *School and Society,* 43: 177–183, February 8, 1936.

[13] J. Wayne Wrightstone, "Measuring Some Major Objectives of the Social Studies,"

Item validity. Specialists in test construction not only attempt to validate the test as a whole against some outside criterion, but also to validate the items on the test individually, usually against an inside criterion, the test as a whole. Undoubtedly an outside criterion would be better, but it is often not available.[15] Although many of the processes for item validation are rather technical and complicated, the essential idea is easy to grasp: The purpose is to determine the difficulty and the discriminating value of each item in the test. Obviously an item missed by everybody or answered correctly by everybody who took the test is of no value in differentiating between good and poor pupils. If the test is for the purpose of determining the extent to which the minimum essentials of a unit or of a course have been mastered, however, the difficulty of the individual items is relatively unimportant and the matter of discrimination is of minor significance. But if the test is to be used over several grades as a basis of classification or school marks, the discriminating value of the items is of major importance. With the exception of a few easy items at the beginning of such a test for the purpose of building morale in the pupils taking it, the items should show a percentage of successes increasing progressively from the poorest pupils to the best. Several studies conclude that it is easier to secure items with a consistent degree of difficulty than with a consistent degree of discrimination.[16]

Only the simpler processes need concern the classroom teacher, for there is considerable doubt whether the elaborate techniques are enough better than the simpler ones to justify the additional labor involved. It is worthy of note that in one study [17] the simple device

---

School Review, 43: 771–779, December, 1935; also J. Wayne Wrightstone, *Appraisal of Experimental High School Practices,* 194 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1936.

[14] Eugene R. Smith, Ralph W. Tyler, and staff, *Appraising and Recording Student Progress,* 550 pages. New York: Harper and Brothers, 1942.

[15] For an excellent discussion of a technique for item validation employing an outside criterion, see Herbert A. Toops, *A Note on Item Selection,* 6 pages. Ohio College Association Bulletin, No. 103, 1936.

[16] *Cf.* Harold D. Carter, "How Reliable Are the Common Measures of Difficulty and Validity of Objective Test Items?" *Journal of Psychology,* 13: 31–39, January, 1942; also R. M. W. Travers, "A Note on the Value of Customary Measures of Item Validity," *Journal of Applied Psychology,* 26: 625–632, October, 1942.

[17] Theodore F. Lentz, Jr., Bertha Hirshstein, and F. H. Finch, "Evaluation of Methods of Evaluating Test Items," *Journal of Educational Psychology,* 23: 344–350, May, 1932.

of comparing the best third and poorest third of the class on each item, and considering those items most valid which showed the highest percentage of successful responses in the highest third as compared with the lowest third, was slightly more effective than the more elaborate methods. Whether one compares the best third with the poorest third, the best fourth with the poorest fourth, or similar proportions of the distribution, seems a matter of small consequence, and the technique is the same.

An illustration will make the procedure clear. Suppose the class has forty pupils, and it is desired to compare the best and the poorest fourths. The ten papers that score the highest are put in one group, and the ten papers that score the lowest are put in the contrasting group. The other twenty papers need not be analyzed, although there is some value in doing so. Next, the number of times each item is answered correctly by the two contrasting groups is determined. Those items which show the greatest percentage of successful responses in the best group above that of the poorest group are best; those which show little or no discrimination in favor of the best group are useless; and those which show a percentage of successes for the best group below that of the poorest group are worse than useless, since they are actually injurious to the validity of the test. It is in the location of these worst items that the procedure is of most value to the classroom teacher.

Frequently, perhaps generally, it will be found that the trouble is in the *wording* of the item, the language being vague, ambiguous, or positively misleading. In that case a rewording of the item may be all that is necessary. At times, however, the difficulty is more obscure, and the item may have to be eliminated altogether. Lindquist found that "adequately" and "advisers" were equally difficult for eighth-grade spellers, but that the former discriminated in favor of the good spellers and the latter in favor of the poor spellers. Difficulty alone, therefore, is not a dependable measure of discrimination, for according to that criterion both items are equally good. Test experts have usually found, however, that the average difficulty of the items in a test is related to the validity of the test as a whole. The rule usually suggested for the construction of tests covering more than one grade is: *Choose items varying in difficulty for the group being measured from those just above zero per cent to those close to 100 per cent, and of such difficulty that the average will be about 50 per cent of the maximum score possible on the test.* Items for a test covering a narrow range discriminate best when their difficulty is such that each item is passed by 50 per cent of the group. Tests designed primarily for instructional purposes, however, may at times be made much easier with good results.

Judging the validity of standard tests. It is always desirable to examine with some care the content of a standard test before deciding to use it. Some of the earlier tests in particular contained serious errors. Upton[18] called attention to some of these in arithmetic tests, and Diamond[19] found 318 errors in 3,303 items making up the content of sixteen widely used tests in biology and general science. Only one test was found to be entirely free from error. A recent study of five tests of English usage found that from 16 to 55 per cent of the items called wrong were actually acceptable according to standards published by the National Council of Teachers of English.[20] Even when there are no errors, the items used often stress the relatively unimportant aspects of the subject.

The test manual also should be examined, because it frequently gives data on the validity of the test; for example, it should tell who made the test, how the items were chosen, what standardization and validation procedure was followed, and other pertinent information. If the author does not give such information to the prospective users, it is safe to assume that the test is of doubtful validity, for it is evident that the author does not attach as much importance to the matter as is desirable.[21] While it is unnecessary to ascribe improper motives to test authors and publishers, most of whom are of a very high type, it is important to recognize that they are nevertheless human, and it is reasonable to make some allowance for a little overenthusiasm about the merits of their own progeny. Whenever available, therefore, the results reported in professional journals by other users are likely to be especially valuable.

In this connection, attention is called to a list of five "critical issues looking toward the improvement of educational measurement," suggested by Ruch:[22]

> 1. There are in use today at least one thousand different educational and mental tests. Convincing critical and statistical data on the validity, reliability, and norms of these measures are available in probably less than 10 per cent of the cases. The publication of such crucial information is an ethical obligation of the test author and publisher. . . .
>
> 2. In view of the situation just mentioned, there is an urgent need for comparative studies of the relative values of existing tests. In most subjects, this need is probably more insistent than the production of new tests. . . .

---

[18] Clifford B. Upton, "The Influence of Standardized Tests on the Curriculum of Arithmetic," *Teachers College Record,* 26: 627–641, April, 1925.

[19] Leon N. Diamond, "Testing the Test-Makers," *School Science and Mathematics,* 32: 490–502, May, 1932.

[20] Karl W. Dykema, "On the Validity of Standardized Tests of English Usage," *School and Society,* 50: 767, December 9, 1939.

[21] *Cf.* G. M. Ruch, "Minimum Essentials in Reporting Data on Standard Tests," *Journal of Educational Research,* 12: 349–358, December, 1925.

[22] G. M. Ruch, "Recent Developments in Statistical Procedures," *Review of Educational Research,* 3: 39–40, February, 1933.

> 3. The reliabilities of all but a few existing tests are far too low for the measurement of individuals, as contrasted with evaluation of groups. . . .
> 4. Mislabelling of tests is the rule rather than the exception in such titles as *diagnostic* and *prognostic*. Very few diagnostic tests show sufficient reliability of total scores for accurate measurement, not to mention the unreliability of the sub-tests individually. . . .
> 5. There is urgent need for a fact-finding organization which will undertake impartial, experimental, and statistical evaluations of tests—validity, reliability, legitimate uses, accuracy of norms, and the like. . . . Independent workers in this field are few as yet, the task is tremendous, and to leave such determinations to authors of tests and publishers is likely only to continue the present chaotic conditions.

In the meantime, while we are waiting for this "impartial, experimental, and statistical evaluation," it should be possible for us to utilize the expert opinion of test makers in the several fields, employing a less elaborate technique perhaps than Kelley used some years ago. An illustration from the life insurance field may prove instructive. An organization wrote the general agents of the leading life insurance companies and asked them to name the ten best companies in their opinion that issued a certain type of policy. As it was to be expected that each agent would place his own company at the head of the list, the ranking of his company was disregarded in the tabulations. The consensus regarding the other companies showed rather remarkable agreement. Something like this for standard tests ought to be considerably more trustworthy than the individual opinion of the author, publisher, or average test user.

A still more scientific attempt to provide the information required by the test user as a basis for an intelligent choice of tests has been made by Buros.[23] In a series of *Mental Measurements Yearbooks* he plans to make available critical evaluation of all recent tests by three or more competent persons independently.[24] These publications will be found indispensable in selecting tests. The reviewers were instructed to make the expected reviews "frankly critical" and "each one was expected to base the appraisals upon his own criteria as to what constituted a good test." The reviewers are described in the "Introduction" as follows:

> In selecting reviewers an effort was made to choose persons representing a wide variety of positions and viewpoints among actual and potential test users. As a result, a very heterogeneous group of reviewers have cooperated

---

[23] Oscar Krisen Buros, *The Nineteen Thirty Eight Mental Measurements Yearbook,* 415 pages. New Brunswick, New Jersey: Rutgers University Press, 1938.

[24] Buros originally planned to publish these volumes biennially, but the schedule, temporarily interrupted by World War II, was resumed in 1947.

> in the preparation of this volume—classroom teachers, city school research workers, clinical psychologists, curriculum specialists, guidance specialists, personnel workers, psychologists, subject-matter specialists, and test technicians. . . . It can be truly said that the reviewers represent no one group or school of thought, unless the reviewers are described as representing all test users—actual and potential—who are considered especially competent in their fields and who have the courage to speak frankly and honestly in appraising a standard test. . . .
>
> Ideally, a reviewer of a standard test, such as a high school Latin test, ought to possess the qualifications of a curriculum and teaching specialist, and a test technician. Unfortunately, all of these qualifications are rarely found in any one person. . . . The average quality of the reviews is likely to be highest when the reviewers discuss only those points which they feel most competent to appraise, even though this practice frequently results in reviews which are not comprehensive. By having three to six persons review each test, the probability of securing a comprehensive appraisal of a test is greatly increased.[25]

The group judgment of even the most competent persons has certain limitations. There remains the troublesome fact that no one test is equally valid for all purposes, or for the same purpose in all situations. Furthermore, there is no way of knowing when a new test may appear with merits so outstanding as to render obsolete earlier tests that have hitherto been entitled to high comparative ratings. But, all things considered, the best available source of information is probably the department of measurement in a reputable college or university, of which there are several in most states. Such recommendations can usually be relied upon to be impartial and based upon a wider acquaintance with existing tests than the average teacher or school administrator is likely to have. But in the final analysis, when all the cards are on the table, the teacher or administrator must rely upon his own judgment. The necessary background for making such a judgment intelligently should be specifically provided for in the professional training of teachers. The data required for such judgments should be made available by the test publishers and by such publications as the *Mental Measurements Yearbooks.*

C. Reliability

Meaning of reliability. By reliability is meant the degree to which the test agrees with itself. To what extent can two or more forms of the test be relied upon to give the same results, or the same test to give the same results when repeated? If the scores on the test are stable under these conditions, the test is said to be reliable. In a word, *reliability means consistency.*[26]

---

[25] Oscar Krisen Buros, *The Nineteen Forty Mental Measurements Yearbook,* pages 12–13.

The terms *reliability* and *validity* are often confused, but there is a clear-cut distinction between them. Reliability, as such, has nothing to do with the truthfulness of the measurement, but is concerned only with its consistency, an entirely different thing. A homely illustration may help to clarify the distinction. A man returns from his vacation with a picturesque story of the fish he claims to have caught. As he meets friend after friend, there is always the same glowing account, even to the minutest detail. Now, in a statistical sense the story is *reliable,* for it is certainly *consistent.* Unfortunately, the fisherman's veracity is not thereby established, for consistency by itself gives no assurance of truthfulness or validity. In reality the story might be sheer fiction from beginning to end.

Importance of reliability. Shakespeare said: "Consistency, thou art a jewel," and he was right. But consistency is not the greatest jewel, whether in a test or elsewhere. By itself consistency, or reliability, is a doubtful virtue, for a test, as well as a person, might be consistently wrong, but its absence is a sign of weakness. Although high reliability is no guarantee that the test is good, low reliability does indicate that it is poor. In the above illustration it should be noted that had marked discrepancies occurred in the fisherman's story from time to time, considerable doubt would have been cast upon his truthfulness. Validity is always the first quality to be sought in a test, and, granted that, reliability is a valuable auxiliary. *The ideal test tells the truth consistently.*

There can be little question that test makers have given too much attention, relatively, to determining the reliability of tests, and too little to establishing their validity. One reason for this, doubtless, is that the former is much easier to determine. Much harm has resulted, however, when uncritical users have naïvely assumed, as pointed out by Monroe, that reliability insures validity, a view which is wholly erroneous.

Methods of determining reliability. The term *reliability* is purely a statistical concept. Contrary to what was found in the case of validity, very little can be told about the reliability of a test from examining the test blank itself. It is, of course, true that if a test can be objectively scored, it is more likely to be reliable than if the scoring is subjective; but the degree of reliability cannot be determined by that fact. It is also true that a long test has a greater likelihood of being reliable than a short test; but there are many exceptions. In the last analysis, however, *somebody must try the test out to determine its reliability.* Usually the author of the test does this and reports the results in the test manual. If such is not the case, one has a right to be suspicious of the merits of the test.

---

[26] For a critical discussion of this concept, see R. W. B. Jackson and G. A. Ferguson, *Studies on the Reliability of Tests,* pages 18–25. Toronto: University of Toronto Press, 1941.

Method with two test forms. Three rather distinct techniques are used to establish the reliability of a test. The method commonly used by makers of standard tests is to prepare two or more parallel forms of the test, and then to give these equivalent forms of the test to a large number of pupils, usually with only a short interval between the tests. The test is said to be reliable if there is close agreement between the scores on the two forms; that is, if the pupils who made high scores on the first test also make high scores on the second; if those who made low scores on the first test again make low scores on the second; and so on for all ranks in between. If the agreement is perfect, as is most unlikely, the correlation is 1.00. On the other hand, if there is no consistent relationship, the coefficient of reliability is .00. It will be recalled that validity is also expressed as a coefficient of correlation whose maximum value is 1.00, and whose minimum value is .00. But in the case of validity the agreement is with an *external* criterion, whereas in the case of reliability the agreement is with an *internal* criterion of some kind. In the above illustration this internal criterion is another form of the same test, which presumably measures the same functions as the first test.

Methods with one test form. When only one form of a test is available its reliability can still be determined. One procedure is to repeat the test at a later time and to determine the extent of agreement by computing the coefficient of correlation between the two series of scores. Another procedure is to give the test once only and then to record two scores for each paper, one for each half. Usually the test is divided into "chance halves" by computing one score for the even-numbered items and another score for the odd-numbered items. When the two series of scores are obtained, the coefficient of correlation between them is computed. This is the reliability of the half test. The reliability of the whole test is then estimated by the use of the Spearman-Brown formula.[27] This same formula also makes it possible to estimate the probable reliability of the test when increased to any required length, assuming that the items added are of the same type and quality as those in the original test.

---

[27] This formula is discussed more fully on pages 244–245.

Both methods for obtaining the reliability of a single form of the test have been severely criticized and as stoutly defended. The test-retest method has certain serious limitations. The most serious limitation is that it mixes two different things in unknown proportions, the variability of the test and the variability of the pupil. If the test is long, to avoid fatigue and boredom, some time must elapse between the two trials. In the case of achievement tests, particularly, this delay is likely to introduce other variables. The pupils may discuss the test between trials, do extra study, or do other things that may effect a change in the status of their knowledge. In addition to this, their physical and mental conditions fluctuate from day to day, even from hour to hour. For example, Ashbaugh [28] found variability in one fourth of the pupils who were given the same spelling test under highly constant conditions three times within fifteen minutes. One would appreciate the difficulty in determining the reliability of a certain type of thermometer by checking the readings made at one hour against those made later in the day. Guilford thinks that "it is safe to say that the average test scale of mental ability is fully as reliable, or probably more so, than the average clinical test in medicine, such as the test of blood pressure or the basal metabolism test, whose reliability ranges from about .60 to .90." [29] But there is also the contrary tendency in human beings for errors made the first time to persist and to be repeated at later times. In an extreme situation, where the pupils memorized the first series of answers, the apparent reliability of the test would be perfect. Indeed, this tendency to echo the original responses appears to be strong, for test-retest coefficients are usually higher than those arrived at by correlating chance halves or equivalent forms of the test. Because the correlation of the half-tests eliminates, or at any rate greatly reduces, the pupil variable, it is recommended by such writers as Anastasi [30] and Jordan.[31]

The chance-half procedure, involving as it does the Spearman-Brown formula, also has certain difficulties. For one thing, the formula is based upon certain assumptions that are rather hard to meet in actual practice, and there is no magic in the formula which enables it automatically to make the necessary adjustments. For example, the average difficulty of the half-tests should be equal, as well as of equal variability, and the items to be added must be of the same quality as those already included. It must be emphasized that the formula requires the use of *chance* halves of the test, not just any halves.[32] It has been empirically demonstrated, however, that when the formula is employed as intended it gives a very close approximation.[33]

---

[28] Ernest J. Ashbaugh, "Variability of Children in Spelling," *School and Society*, 9: 93–98, January 18, 1919.

[29] J. P. Guilford, "Intelligence Tests," *Education*, 58: 528, May, 1938.

[30] Anne Anastasi, "The Influence of Practice Upon Test Reliability," *Journal of Educational Psychology*, 25: 321–335, May, 1934.

[31] R. C. Jordan, "An Empirical Study of the Reliability Coefficient," *Journal of Educational Psychology*, 26: 416–426, September, 1935.

Kuder and Richardson [34] have devised a simpler method of obtaining a reliability coefficient which makes it unnecessary to split the test into halves or calculate a coefficient of correlation. Unfortunately, the procedure involves certain assumptions which are likely to be difficult to meet in the ordinary test situation.[35]

Goodenough [36] concludes that "there is not any best method" of computing reliability coefficients, and suggests that the purpose of the study should determine the most appropriate methods to use for each situation. The possibility of locating discrepancies where more than one method is employed is also pointed out. The data available abundantly establish the fact that no two methods are likely to produce identical results in any given situation. The obligation to indicate the particular method used is evident.

The interpretation of test reliability. What standard shall a test meet in order to be considered satisfactory from the standpoint of reliability? No simple answer to this question is possible. It depends, for one thing, upon the fineness of discrimination required. Kelley [37] has suggested the following minimal requirements for the reliability coefficients of a single school grade:

.50 for determining the status of a group in some subject or group of subjects.

.90 for differentiating the achievement of a group in two or more scholastic lines.

.94 for differentiating the status of individuals in the same subject or group of subjects.

.98 for differentiating individuals in two or more scholastic lines.

The interpretation is also beset by many other difficulties. The coefficients not only reflect somewhat the methods employed in their computation, but also the variability of the groups, the interval between tests, and other factors. For example, a test of average difficulty for a normal group appears to be much less reliable when used with a markedly inferior or markedly superior group. To escape some of these factors, Ruch,[38] Odell,[39] and others have suggested that the reliability of the test be expressed in some unit which can be represented as a ratio to the mean or standard deviation. Units which meet this requirement are the standard and probable errors of measurement or the standard and probable errors of estimate.

---

[32] *Cf.* William A. Brownell, "On the Accuracy with Which Reliability May Be Measured by Correlating Test Halves," *Journal of Experimental Education,* 1: 204–215, March, 1933.

[33] Giles M. Ruch, Luton Ackerson, and Jesse D. Jackson, "An Empirical Study of the Spearman-Brown Formula as Applied to Educational Test Material," *Journal of Educational Psychology,* 17: 309–313, May, 1926.

[34] G. F. Kuder and M. W. Richardson, "The Theory of the Estimation of Test Reliability," *Psychometrika,* 2: 151–160, September, 1937.

[35] Truman L. Kelley, "The Reliability Coefficient," *Psychometrika,* 7: 75–83, June, 1942.

[36] Florence L. Goodenough, "A Critical Note on the Use of the Term 'Reliability' in Mental Measurement," *Journal of Educational Psychology,* 27: 173–178, March, 1936.

[37] Truman Lee Kelley, *Interpretation of Educational Measurements,* pages 28–29. Yonkers: World Book Company, 1927.

But even this procedure does not solve all the problems. Lindquist and Cook,[40] for example, found that both the reliability and validity of self-administering spelling tests "are to a very significant degree a function of the time in which they are administered." There was a general tendency for validity to increase with the increase in testing time and for the reliability to decrease up to a certain length of period, beyond which both reliability and validity remained constant. It appears, therefore, that increases in reliability can be bought at too dear a cost. An English psychologist[41] points out that when tests are used in batteries, "reliability may or may not be a desirable feature of a test." In fact he contends that low reliability in a test "may be actually a virtue in some circumstances." His major conclusion is that the "factor saturation, rather than the reliability coefficient, gives the important information required in regard to a test." In this connection, it may be observed that such "factor saturation" has so far been determined for no standard tests in general use.

In view of the fact that measures of reliability, no matter how arrived at, are influenced by factors other than the form and content of the test itself, it would appear that the value of such measures has been overemphasized.[42] The same energy devoted to improving the validity of the test would bring better returns. It is not likely that the average teacher will find it profitable to compute reliability coefficients for ordinary class tests, although it may be worth while to do so for final examinations.

---

[38] G. M. Ruch, "Minimum Essentials in Reporting Data on Standard Tests," *Journal of Educational Research,* 12: 349–358, December, 1925.

[39] C. W. Odell, *Educational Measurement in High School,* pages 63–66. New York: D. Appleton-Century Company, 1930.

[40] E. F. Lindquist and Walter W. Cook, "Experimental Procedures in Test Evaluation," *Journal of Experimental Education,* 1: 163–185, March, 1933.

[41] W. Stephenson, "Factorizing the Reliability Coefficient," *British Journal of Psychology,* 25: 211–216, October, 1934.

[42] Some writers would abandon altogether the "blanket term reliability" in favor of more specific estimates of absolute and relative accuracy of measurement. *Cf.* Jackson and Ferguson, *op. cit.,* page 25.

Objectivity and reliability. By objectivity in a measuring instrument is meant the degree to which equally competent users get the same results. Ordinary measures of height and weight, for example, are objective, while estimates of beauty and integrity are subjective. The distinction between objective measurement and subjective measurement is implied in the question: "Do married men *really* live longer than single men, or does it just *seem* longer?" As a rule, objectivity is very closely associated with reliability. For this reason standard tests are usually more reliable than rating scales. As a matter of fact, great impetus was imparted to the objective test movement by the discovery that the major cause of the notorious unreliability of the ordinary school examination was its subjectivity of marking. The emphasis on objectivity has since gone so far, however, that many educational workers seem to regard "objectivity" as synonymous with "scientific method." To such persons, any element of subjectivity in a study renders it hopelessly unscientific. It may be well, therefore, to look carefully at this all-important matter of objectivity.

To discover at the outset that there is no such thing as a *wholly objective* measure may be something of a shock. The plain fact is that objectivity is always relative, never absolute. The measurements obtained by a yardstick, for example, are only relatively objective, for one would hardly expect a dozen different persons to get absolutely the same results in measuring the length of the playground. They would probably agree to the nearest foot, and possibly to the nearest inch, but they would usually disagree markedly, if the results were expressed in some such small unit as hundredths of an inch. And, of course, such units as inch, foot, and yard are not *natural* units, like day and year, but units set up by human judgment.

Brownell[43] points out that there are always many subjective factors involved even when the test used is of the so-called objective type. He says:

> Well, first of all, in the practical circumstances of teaching, one *decides to give* a test. The decision is surely not based upon purely objective considerations. Second, one determines whether to *make* a test or to *buy* one. . . . Third, one makes up one's mind regarding the *kind* of test—whether it is to be of the traditional type, of the newer types, or a combination—judgment again. Fourth, one settles upon the *scope* of the test—judgment once more. Fifth, one selects the *items* to be included—little objectivity here. Sixth, one chooses the *form* to be employed—true-false, multiple choice, or what not—again little objectivity. Seventh, one *frames the items* as carefully as one can—

---

[43] William A. Brownell, "The Use of Objective Measures in Evaluating Instruction," *Educational Method,* 13: 401–408, May–June, 1934.

and once more has only his judgment for guidance. Eighth, one prepares a *key* by listing the correct answers—a judgment which may not be acceptable to other teachers even of the same subject. Ninth, through opinion one defines the conditions of *administering* the test. Tenth, one *scores* the papers—at last objectivity. But, eleventh, one *assigns marks*—another increment of judgment, and a big one.

Brownell protests against what he regards as the overemphasis on objectivity, which he thinks has unnecessarily lessened the depth and narrowed the range of measurement. A safe position would appear to be to *try to make measurement as objective as possible without sacrificing validity.* It must be remembered that the latter is always more important. It is never going to be possible or desirable to eliminate certain basic assumptions underlying all attempts at evaluation. Undoubtedly at times, however, we have made *assumptions* in measurement when we should have had *evidence.* Many test makers, for example, have assumed that one problem of a type is sufficient for diagnosis in arithmetic. When the matter was actually subjected to experimental analysis in two studies,[44] both found that one problem of a type is likely to be both unreliable and invalid, owing largely to chance, and that at least three problems of each type must be included for satisfactory individual diagnosis. Another assumption which did not check with the evidence was that objectivity of scoring guarantees accuracy of scoring. Several studies have demonstrated the fact that scorers of standardized tests must be *taught* and not merely *told* how to do it. Dearborn and Smith,[45] for example, found that 73 per cent of all tests investigated contained errors, three fourths of which were in the direction of too severe marking. A serious effort should of course be made to eliminate all needless types of subjectivity. A guess is usually a poor substitute for actual knowledge.

D. Usability

Meaning of usability. There is quite general agreement among authorities in measurement that the two most important characteristics of a measuring instrument are validity and reliability. Both have to do with the theoretical accuracy with which the instrument measures. However there are certain other considerations of a practical character which must be taken into account. In the judgment of the writer all of these may be conveniently designated by the single term *usability.* By this is meant the degree to which the test or other instrument can be successfully employed by classroom teachers and school administrators without an undue expenditure of time and energy—in a word, *usability means practicability.* A measuring instrument must not only be valid and reliable but also usable. This viewpoint is well expressed in *The Methodolgy of Educational Research:* [46]

---

[44] Leo J. Brueckner and Mary Elwell, "Reliability of Diagnosis of Error in Multiplication of Fractions," *Journal of Educational Research,* 26: 175–185, November, 1932; Foster E. Grossnickle, "Reliability of Diagnosis of Certain Types of Error in Long Division with a One-Figure Divisor," *Journal of Experimental Education,* 4: 7–16, September, 1935.

[45] Walter F. Dearborn and C. Wilson Smith, "The Results of Rescoring Five Hundred Thirty Dearborn Tests," *Journal of Educational Psychology,* 20: 177–183, March, 1929.

> But we must always temporize ideals with practical considerations. Perhaps an *ideal* instrument would be so cumbersome and expensive of effort and time that its use would not be warranted.

Whether or not a test is usable by average teachers in service and other persons whose technical training in measurement has been limited depends upon several factors, of which the following are probably the most important:

1. Ease of administration.
2. Ease of scoring.
3. Ease of interpretation and application.
4. Low cost.
5. Proper mechanical make-up.

Each of these factors will now receive brief consideration.

Ease of administration. Group tests, as a rule, are much easier to administer than individual tests. The Stanford-Binet is a good example of a test whose validity and reliability are high, but whose usability is low, largely because of complicated instructions for giving and scoring. Special training in a college course for one semester is usually suggested as the minimum required for mastery of these instructions. Even then the test makes heavy demands upon the examiner's time.

There are, of course, two types of instructions for a test. One has to do with directions to the examiner, and the other has to do with directions to the pupil or pupils. But, in general, the requirements are the same for both. The motto of a well-known news weekly indicates what is required: The directions should be "clear, curt, complete." Whether or not examples, fore-exercises, and the like are necessary will depend mainly upon the age and experience of the group being examined. Whether or not a group test is easy to administer depends to a considerable extent upon the completeness of the manual. Some tests have no time limits, many have generous time limits, while still others are broken up into intervals as short as 3, 5, 8, 10, or 15 seconds. These short intervals are difficult to observe with a stop watch and well-nigh impossible without it. On the other hand, tests of the so-called self-administering type involve only one short set of directions for the entire test. Most tests, however, are broken up into separate sections, each of which has its own directions and time limit. In determining how difficult a test is going to be to administer, a careful examination must be made both of the manual and of the test blank itself.

---

[46] Carter V. Good, A. S. Barr, and Douglas E. Scates, *The Methodology of Educational Research,* page 439. New York: D. Appleton-Century Company, 1936.

Ease of scoring. The ease of scoring a test depends primarily upon three things: objectivity, adequate keys, and full scoring directions. The better standard tests rank high on all three counts. Scoring is also facilitated when the pupil has been instructed to record his answers in a straight column rather than irregularly over the page, and in the form of a numeral or single word rather than a phrase or longer statement. As a rule, all acceptable answers should appear on the key. With the exception of rating scales in which score values of unequal weight are required, all items should be weighted equally. The unequal weighting of items, so common in the earlier tests, has been found to add to the difficulty of scoring without a corresponding increase in validity or reliability.

But even when all these conditions are met, a considerable amount of time is required in scoring. In the past few years much ingenuity has been shown in devising ways of cutting down this time requirement. Most of the methods suggested involve the use of answer sheets of some kind. Several years ago Ross and Gard[47] found that standard tests could be given by the answer-sheet method, or even dictated to pupils, without materially reducing either reliability or validity. Toops has devised intelligence tests[48] with separate answer sheets in triplicate, on which the student records his answers by punching holes with a stylus. The paper is then scored by a machine, and three complete records are at once available for various school officials. Pressey[49] has invented special machines for automatically recording the score when the student presses a key, which is the method of indicating his response to an item. The Clapp-Young self-marking tests[50] and answer sheets have been commercially available for years. Cuff[51] has designed a clever machine in which a series of metal rods are so arranged as to drop through when the pupil has answered correctly, and are literally weighed on a pair of commercial scales. Machines for mechanically recording scores by electrical contacts are available commercially.[52]

---

[47] C. C. Ross and Paul D. Gard, *Two Modified Methods of Administering Two Group Intelligence Tests,* 115 pages. Bulletin of the Bureau of School Service, University of Kentucky, Vol. 2, June, 1930.

[48] Published by Ohio State University.

[49] See *School and Society,* 23: 373–376, March 20, 1926; 25: 549–552, May 7, 1927; 36: 668–672, November 19, 1932.

[50] Published by Houghton Mifflin Company.

The tremendous advantage of these improved methods is illustrated by the experience of Stenquist,[53] director of research in Baltimore. He reports that a test was given to one of the largest high schools in the city; the answer sheets were delivered to his department at three o'clock that afternoon, were scored in one and a half hours, and were returned to the school the following morning ready for use. Today it is no more necessary laboriously to score tests by hand than it is to print them by hand. Also, where quantity production is required, machine methods reduce the cost.

Ease of interpretation and application. Whether or not the results of a test are easy to interpret and apply depends primarily upon the adequacy of the manual accompanying the test. In the first place, the manual should contain complete norms to facilitate interpretation. Whenever possible, all derived scores should be capable of being read directly from tables of norms without the necessity of computation. The norms should, as a rule, be based both on age and on grade; and, in the case of high-school achievement tests, on the length of time the subject has been studied. It is also desirable that all achievement tests should be provided with separate norms for city, town, and rural pupils, and for pupils of various degrees of mentality. Up to the present time very few tests are adequately provided with norms for interpretation. Where the primary emphasis is upon diagnosis and other instructional values of tests, this loss is not very great. In any event, it will always be necessary to rely heavily upon local norms.[54]

Several of the better manuals give specific suggestions regarding the use to be made of the test results. The *Supervisor's Manual* accompanying the Metropolitan Achievement Tests[55] is a good example. Supplying some suggestions as to results is a valuable service for which it is hoped test publishers in the future will accept more responsibility. For many uses it is necessary to have at least two forms of the test equated both as to content and as to difficulty throughout the full range of scores, and not averages only. Very few tests meet this requirement fully. A critical summary[56] states:

---

[51] Noel B. Cuff, "A New Device That Scores Tests," *Journal of Educational Psychology,* 26: 73–77, January, 1935.

[52] For a comprehensive summary and bibliography on test-scoring by machine methods, see Irving Lorge's discussion in *Review of Educational Research,* 12: 550–557, December, 1942.

[53] John L. Stenquist, "Experiments with Machine Scoring of Tests," pages 83–85. *Baltimore Bulletin of Education,* Vol. 13, September–November, 1935.

[54] A fuller discussion of norms will appear in Chapter X.

[55] Published by World Book Company.

> . . . With two or three outstanding exceptions, most of the so-called "standardized" tests available up to the year 1932 have existed in only two "equivalent" forms, which in some instances, at least, have turned out to be only "approximately" equivalent. We have had several series of so-called intelligence tests and several series of achievement tests in each of several matters for ten years or more; but even yet it may be confidently asserted that, with minor exceptions, no two such series have been made comparable, even though they have been edited by the same editor and published by the same publishing house.

The list of "outstanding exceptions" has been extended somewhat since 1932, but in the main the indictment still holds. It is possible that the foregoing standard sets an ideal which can be only approximated, never fully attained. Nevertheless, the lack of equivalent tests imposes a severe restriction upon many forms of educational research. For example, Watson[57] and others have called attention to the fact that many tests are not suitably scaled for measuring growth. For measuring progress over short periods of time the test must be scaled into fine units; otherwise it will be impossible to detect improvement even when it is present. The ordinary "health" scales may be sufficiently accurate for mother's weight, but hardly sufficient for detecting baby's growth from day to day.

Cost. With the exception of certain laboratory apparatus and equipment for measuring special abilities and disabilities, testing materials are usually not very expensive. Few achievement tests covering a single subject, or group tests of general intelligence, cost more than five cents each. Batteries covering several subjects when printed as a single booklet usually cost from five to ten cents. For a comprehensive testing program the general battery will cost less than separate tests covering the same subjects. Cost is a practical consideration in most school systems, and there is no point in paying more for tests than necessary.

While it may be true in general, as commonly held, that in the long run one gets about what he pays for, there are too many exceptions to make it a safe rule. In statistical terminology the correlation between the cost of a test and its worth is positive, but too low for accurate prediction. Here, as elsewhere, the customer should be wary, lest he not get his money's worth. In a test, as in an automobile, the quality is often not evident on the surface. The prospective purchaser should not make cost a primary consideration, for good tests are often no more expensive than poor ones. Fortunately, therefore, relative cost can be considered a minor matter, as a rule, and the choice of the test can rest, as it should, upon its validity and reliability for the purpose it is to serve. It must be remembered that one test may be cheap enough at five cents and another too costly at one cent. After all, the careful purchaser is more concerned with what he gets for his money than with what he has to pay.

---

[56] Ben D. Wood, E. F. Lindquist, and H. R. Anderson, "Basic Considerations," *Review of Educational Research,* 3: 5, February, 1933.

[57] Goodwin Watson, "Note on Validity in the Measurement of Change," *Journal of Educational Research,* 27: 187–192, November, 1933.

Mechanical make-up of the test. Tests issued by the larger publishers are almost always printed in clear type of a size appropriate to the grade level for which they are intended. But there are some exceptions. Not long ago one of the leading publishers issued a test in which the key word in each sentence was supposed to be in bold-faced type, but in many cases the quality of the type did not clearly indicate which word was intended. On a timed test such as this one, a handicap was imposed upon all pupils except those with the keenest vision. In the lower grades careful attention should be given to the quality of pictures and illustrations used. In the earlier days it was common to have the instructions to the examiner appear on the test booklet in the hands of the pupil. This practice not only meant a needless cost in paper and printing, but was a possible source of confusion to the pupil.

Commercial publishers of tests have not as yet given sufficient attention to devising tests which will reduce to the minimum their cost in time and money. There appears to be no valid educational reason why tests should not be designed with separate answer sheets, a practice which would not only eliminate the economic waste of using the test one time only and then discarding it, but which would also facilitate greatly the scoring and make the pupil's test profile available in convenient form for use and filing. It is likely that when test users insist on these improvements they will get them. Absolutely nothing is gained, however, when the answer sheet itself is sold at prices almost as high as the test itself, as is now often the case. The customer usually gets what he wants when he wants it badly enough and makes his wishes known. To make convenient, inexpensive tests feasible, moreover, the demand must be sufficiently increased so that the additional volume sold compensates for reduced profit per unit.

Summary. What, then, are the earmarks of a good measuring instrument? In brief, a good test or other measuring instrument possesses three outstanding qualities: validity, reliability, and usability. In other words, a *good test measures what it claims to, consistently, and with a minimum expenditure of time, energy, and money.* But always the first consideration is validity. The test must not only measure what it purports to, but in the case of achievement tests, it should purport to measure the really important outcomes of the educational process. No test that fails to do this can be considered a satisfactory measuring instrument, whether made by the classroom teacher or purchased from a publisher of standard tests.

E. Some Generalizations Regarding the Problem of Measurement

The role of measurement in science in general and in education in particular, has been set forth in the first chapter; the historical development of measurement in education has been traced in the second chapter; and the characteristics of a satisfactory measuring instrument have been described in the present chapter. In the light of these discussions a few important generalizations will now be attempted.

1. *Some kind of measurement or evaluation is inevitable in education.* This generalization is amply supported by the history of every recognized science, and of education itself, regardless of whether it is to be classified as a full-fledged science or not.

2. *All measurement is subject to error.* This is true of the so-called "exact sciences"; and to a greater degree it is true of the less exact or newer social sciences, such as psychology and education. Westaway, for example, thinking mainly of physics and chemistry, concludes: "We may, in fact, look upon the existence of error in all measurements as the normal state of things."[58] Kelley speaks of the "ubiquitous probable error"[59] in psychology and education. These errors can be reduced but never wholly eliminated.

3. *These errors of measurement are due in part to the imperfection in the measuring instruments available.* There are no perfect measuring instruments, even in the physical sciences. Westaway, for example, says that "even the very best of the instruments with which we perform our measurements are imperfect."[60] This is true of the fundamental units of measurement in the physical sciences, as well as of the biological and social sciences. No astronomer knows precisely the velocity of light, and yet the light year is the yardstick of celestial measurement; no chemist knows the pre-

---

[58] F. W. Westaway, *Scientific Method: Its Philosophical Basis and its Modes of Application,* pages 289–290. New York: Hillman-Curl, Inc., 1937.

[59] Truman Lee Kelley, *Interpretation of Educational Measurements, op. cit.,* page 19.

[60] F. W. Westaway, *op. cit.,* page 286.

cise value of a single atomic weight, and yet it is the basic unit in chemical analysis. In psychology and education these imperfections are an even more potent source of error than in the older sciences. However it must be remembered that the tools of measurement are much better than they used to be. Summarizing the situation for intelligence tests, Thorndike says: [61]

> Existing instruments represent enormous improvements over what was available twenty years ago, but three fundamental defects remain. Just what they measure is not known; how far it is proper to add, subtract, multiply, divide, and compute ratios with the measures obtained is not known; just what the measures obtained signify concerning intellect is not known.

4. *The limitations of the methods used are a still more important source of error in measurement.* Again this difficulty is true of the physical sciences as well as of the social sciences. For example, Max Planck says that in physics "every measurement, however exact, inevitably involves certain errors of observation." [62] These errors are due partly to sensory and temperamental defects, and partly to lack of skill in the observer. But a still more troublesome source of error is the tendency for the act of observation to interfere with the phenomena being observed in measurement. Heisenberg, for example, noted that the "measurement of an electron's velocity is inaccurate in proportion as the measurement of its position in space is accurate, and vice versa," [63] owing to the disturbing influence of the light rays falling on it in the act of measurement. From this discovery resulted the famous "uncertainty principle" or the "principle of indeterminacy," which has profoundly influenced modern physics. "As a matter of fact every measurement," says Planck, "whatever the method of its employment, invariably interferes more or less with the event to be measured." [64] But this interference is so slight as to be of theoretical interest only to the laboratory physicist engaged in the study of aggregates of elements instead of individual electrons. And the ordinary Newtonian principles of chemistry and physics still operate in the usual way in such practical realms as engineering and medicine.[65]

But the disturbing effect of the measurement process is more serious in education. The personality of the examiner, as well as the testing materials, is always part of the test situation.[66] This is recognized in giving individual tests, where a proper *rapport* between examiner and subject is regarded as essential to a successful examination. But here, even with skilled examiners, the factor is rarely eliminated altogether, for it has been found that the IQ remains more stable when the same examiner gives all the tests. In all experiments, whether involving the use of individual or group tests, the subjects are not purely naïve and receptive creatures but are actuated by motives of pride, desire to please or make a good impression on the examiner, and the like. In other words, the examiner or experimenter is an important part of the situation, and it is doubtful whether standardized instructions can ever reduce this part to the point at which it is negligible. The factor is especially important in character and personality measurement and in the evaluation of social behavior. Certainly one would hardly expect to get as normal reactions of love-making in the psychological laboratory during the day as he would if he were concealed in a tree beside a bench in the park during the evening.

---

[61] Edward L. Thorndike and others, *The Measurement of Intelligence,* page 1. New York: Bureau of Publications, Teachers College, Columbia University, 1927.

[62] Max Planck, *The Philosophy of Physics,* page 24. New York: W. W. Norton & Company, Inc., 1936.

[63] *Ibid.,* page 62.

[64] *Ibid.,* page 68.

[65] For a stimulating discussion of the practical implications of this uncertainty principle by a distinguished American chemist, see: Irving Langmuir, "Science, Common Sense, and Decency," *Science News Letter,* 43: 3–4, 12–15, January 2, 1943.

The principal limitation of measurement due to the human factor in the equation is well stated by Barr: [67]

> In general, however, the good and evil of measurement can be attributed to either (a) the adequacy or inadequacy of the persons who make the measurements: their abilities, knowledges, skills, attitudes, ideals and interests in measurement, or (b) the adequacy or inadequacy of the instruments of measurement: their validity, reliability, objectivity, etc. While our instruments for measuring many of the products of instruction are wholly inadequate and while it will probably take years of painstaking effort to develop adequate instruments for the measurement of these more difficult to measure products of instruction, the chief difficulty today in spite of all this, lies it seems, not so much in the inadequacy of our instruments of measurement as in the inadequacy of the persons who use them: their failure to choose measurements in terms of the specific purpose to be served; their use of instruments of measurement for purposes for which they were never intended; the neglect or nonmeasurement of products of instruction that cannot be measured objectively; their treatment of reliable instruments as if they were valid; etc.

5. *Teachers and school administrators must not only understand and appreciate the functions of measurement in education, but they must realize more fully the limitations of present measuring instruments.* In the present state of measurement two erroneous attitudes are sometimes found. The first is that held by certain overenthusiastic supporters of measurement, who make unreasonable

---

[66] See Saul Rosenzweig, "The Experimental Situation as a Psychological Problem," *Psychological Review*. 40: 337–354, July, 1933.

[67] Clifford Woody and others, "A Symposium on the Effects of Measurement on Instruction," *Journal of Educational Research*, 28: 482, March, 1935.

claims for existing measuring instruments, and who gloss over or refuse to recognize the imperfections that exist.[68] This attitude is not unlike that of the adolescent in his first love affair, where, indeed, if love is not actually blind, it deliberately closes its eyes; and in any event the result is the same. This point of view is unfortunate and unintelligent, for it stands in the way of progress toward needed improvements; making such unwarranted claims is the surest way to discredit the movement with thoughtful people.[69] Fortunately, this attitude appears to be on the decline.[70]

But a second and equally erroneous attitude goes to the opposite extreme. It characterizes those who are as blind to the virtues of existing measuring instruments as the first group are to their limitations, and who refuse to have anything at all to do with tests and examinations until all defects are forever removed. This attitude is as unreasonable as that of the farmer who has postponed buying an automobile "till them blamed things is perfected," and who has in the meantime worn out a great deal of shoe leather without seeing much of the world either.

Then there is the third attitude, that of the practical person who has learned through experience not to expect perfection. Moreover, he has found that excellent work can often be turned out with imperfect tools, if only they are used with sufficient skill. He has also discovered that greater skill is called for than if the instruments were perfect, and he sets out deliberately to attain the skill needed. He realizes that the very existence of these imperfections imposes a special obligation upon the user to seek to understand as fully as possible their nature in order to get desired results in spite of them. Furthermore, he makes a conscious effort in interpreting and using test results in order to take into account the existence of errors. In other words, he takes the very common-sense point of view that the proper thing to be done under the circumstances is to make the best possible use of such tools as exist, while waiting for better ones to be developed.

---

[68] For a suggestive analytical discussion of the problem, see: Douglas E. Scates, "Differences Between Measurement Criteria of Pure Scientists and of Classroom Teachers," *Journal of Educational Research,* 37: 1–13, September, 1943.

[69] The author recalls having heard a gray-haired southern educator say, regarding intelligence tests, soon after World War I: "The worst enemies of any new cause are its darn fool friends!"

[70] Picturesque pleas for sanity in using tests by two pioneers in test-construction are made in S. A. Courtis' "Let's Stop This Worship of Tests and Scales," *Nation's Schools,* 31: 16–17, March, 1943; and in Guy M. Wilson's "Some Subversive Activities of the Test Expert," *Educational Method,* 21: 342–343, April, 1942.

SELECTED REFERENCES FOR FURTHER READING

Freeman, Frank N., *Mental Tests, Their History, Principles and Applications* (Revised Edition). Boston: Houghton Mifflin Company, 1939. Chapters IX–XI.

Greene, Harry A., Jorgensen, Albert N., and Gerberich, J. Raymond, *Measurement and Evaluation in the Secondary School.* Boston: Longmans, Green and Company, 1943. Chapter IV.

Guilford, J. P., *Fundamental Statistics in Psychology and Education.* New York: McGraw-Hill Book Company, 1942. Chapter XIV.

Hawkes, Herbert E., Lindquist, E. F., and Mann, C. R., *The Construction and Use of Achievement Examinations.* Boston: Houghton Mifflin Company, 1936. Part I.

Hull, Clark L., *Aptitude Testing.* Yonkers: World Book Company, 1928. 535 pages.

Jackson, Robert W. B., and Ferguson, George A., *Studies on the Reliability of Tests.* Toronto: Department of Educational Research, University of Toronto, 1941. 132 pages.

Kelley, Truman Lee, *Interpretation of Educational Measurements.* Yonkers: World Book Company, 1927. Chapters II–IV.

McCall, William A., *Measurement.* New York: The Macmillan Company, 1939. Book Two.

McNemar, Quinn, *The Revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin Company, 1942. 187 pages.

Odell, C. W., *Educational Measurement in High School.* New York: D. Appleton-Century Company, 1930. Chapter III.

Pintner, Rudolph, *Intelligence Testing, Methods and Results* (New Edition). New York: Henry Holt & Company, 1931. Chapters IV–VII.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman & Company, 1929. Chapter II.

Saucier, W. A., *Introduction to Modern Views of Education.* Boston: Ginn and Company, 1937. Chapters XVII and XVIII.

Smith, B. Othanel, *Logical Aspects of Educational Measurement.* New York: Columbia University Press, 1938. 182 pages.

Stoddard, George D., *The Meaning of Intelligence.* New York: The Macmillan Company, 1943. Parts I and II.

Terman, Lewis M., and Merrill, Maud A., *Measuring Intelligence.* Boston: Houghton Mifflin Company, 1937. Part I.

Wrightstone, J. Wayne, *Appraisal of Experimental High School Practices.* New York: Bureau of Publications, Teachers College, Columbia University, 1936. 194 pages.

PART II

THE CONSTRUCTION OF INFORMAL TEACHER-MADE TESTS

CHAPTER IV

General Principles of Test Construction

Importance of the problem. There are at least three reasons why the development of proficiency in constructing informal teacher-made tests is important. In the first place, the vast majority of tests in use by classroom teachers are of this type. Table 5, based upon a nation-wide study of 1,600 high-school teachers,[1] shows that the typical teacher during one semester uses 7.6 teacher-made tests, 12.6 short quizzes, but fewer than one standardized achievement test (since 60 per cent give no standardized tests at all). In the second place, both traditional examinations made and marked by untrained teachers and new-type tests used by ordinary classroom teachers produce highly unsatisfactory results. The extensive literature on traditional examinations, briefly summarized in

TABLE 5

FREQUENCY WITH WHICH HIGH-SCHOOL TEACHERS GIVE TEACHER-MADE TESTS, SHORT QUIZZES, AND STANDARDIZED ACHIEVEMENT TESTS DURING ONE SEMESTER (AFTER LEE AND SEGEL)

DEPARTMENT	MEDIAN NUMBER OF TEACHER-MADE TESTS GIVEN	MEDIAN NUMBER OF SHORT QUIZZES GIVEN	PER CENT OF TEACHERS GIVING NO STANDARDIZED ACHIEVEMENT TESTS
Mathematics	10.5	18.2	50
Science	9.7	11.4	64
Commercial studies	9.6	13.4	44
Social studies	8.3	13.3	71
Latin	8.0	28.5	61
English	7.0	13.0	46
Foreign language	6.9	19.6	59
Home economics	6.0	7.1	83
Industrial arts	4.6	6.5	78
Physical education	4.0	.9	83
Fine arts	3.4	4.2	88
Total Group	7.6	12.6	60

---

[1] J. Murray Lee and David Segel, *Testing Practices of High-School Teachers,* pages 2–5. United States Office of Education Bulletin, No. 9, 1936.

Chapter II, has repeatedly demonstrated this fact. Evidence[2] is also available to show that amateurs may at times do even worse with the new-type tests than with traditional examinations. For example, Pullias[3] found somewhat greater variability between 68 pairs of short objective tests made by two teachers to cover the same material than has been reported for essay examinations. Incredible as it may be, it does seem possible, although certainly not necessary, to make new-type tests of lower reliability than the traditional examinations. In the third place, both logical considerations and statistical analyses indicate that skillfully prepared informal tests are as reliable and as valid as available standardized tests.[4] In fact, where the teaching conditions are unusual, or where the subject matter is not thoroughly stabilized, as in civics and modern history, such tests may be even more valid. A state-wide survey of high-school achievement conducted in Tennessee,[5] for example, showed that only 56 per cent of the questions in the standardized social studies test then in use could be answered from the state-adopted textbook.

This chapter will consider the general principles of constructing informal teacher-made tests grouped under the following four headings, which indicate roughly the steps or stages in the process:

1. Planning the test.
2. Preparing the test.
3. Trying out the test.
4. Evaluating the test.

A. Planning the Test

It should be recognized at the outset that the construction of satisfactory measuring instruments is one of the most difficult duties the teacher has to perform. Good tests do not just happen. Nor are they the result of a few moments of high inspiration or exaltation. On the contrary the process is calm, deliberate, and time-consuming. Perhaps the best that can be hoped for under existing conditions is that the teacher prepare reasonably comprehensive and adequate informal tests in one subject each year. Best results will usually be obtained from cooperative effort. The procedure employed in developing the Cooperative Achievement Tests, outlined on page 72, is a good illustration. Another example is the plan used by the General College of the University of Minnesota in constructing achievement examinations.[6] Perhaps the best illustration is the procedure followed by the Evaluation Staff of the Eight-Year Study sponsored by the Progressive Education Association.[7] The six major steps in the process as set forth in detail by Smith and Tyler[8] may be summarized briefly as follows:

---

[2] For an early study, see C. C. Crawford and D. A. Raynaldo, "Some Experimental Comparisons of True-False Tests and Traditional Examinations," *School Review,* 33: 698–706, November, 1925.

For a general study, see L. L. Thurstone, "An Appraisal of the Test Movement," in *Tests and Measurements in Higher Education,* edited by William S. Gray, pages 128–137. Chicago: University of Chicago Press, 1936.

[3] Earl V. Pullias, *Variability in Results from New-Type Achievement Tests,* pages 46–47. Durham, North Carolina: Duke University Press, 1937.

[4] Henry Daniel Rinsland, *Constructing Tests and Grading in Elementary and High School Subjects,* pages 296–298. New York: Prentice-Hall, Inc., 1938.

[5] Jos. E. Avent, *Report of the Tennessee State Testing Program,* page 83. Nashville: State Department of Education, 1946.

1. The faculty of each school was asked to formulate a careful statement of its educational objectives.
2. Statements from these thirty schools were classified by the Evaluation Staff into ten major types of objectives.
3. Each type of objective was then defined in terms of expected pupil behavior.
4. Situations were suggested in which pupils could be expected to show the particular kind of behavior.
5. The more promising methods of obtaining evidence regarding each type of objective were then selected from existing techniques or devised by the staff, and subjected to experimental trial.
6. The methods which made the best showing in this preliminary trial were further developed and improved.
7. Means were devised for the interpretation and effective use of the various instruments of evaluation.

It is recognized that the process just described is too elaborate for the ordinary school, or for the individual teacher working on his own. However, it cannot be emphasized too strongly that the actual process of test construction must be preceded by careful planning if the test is to succeed. The test will be no better than the quality of the thinking that goes into it. In planning the test, consideration must be given to the nature of the objective to be measured, the purpose it is to serve, and the conditions under which it will be used.

1. *Adequate provision should be made for evaluating all the important outcomes of instruction.* A careful statement of the philosophy of the school and the objectives of the particular course should be available from the start. A survey[9] of a representative

---

[6] Alvin C. Eurich, "Evaluation of General Education in College," *Journal of Educational Research,* 35: 502–516, March, 1942.

[7] Published in five volumes by Harper & Brothers, New York, 1942, under the general title *Adventure in American Education.*

[8] Eugene R. Smith, Ralph W. Tyler and Evaluation Staff, *Appraising and Recording Student Progress,* Chapter I. New York: Harper & Brothers, 1942.

[9] B. E. Leary, *A Survey of Courses of Study and Other Curriculum Materials Published Since 1934.* Washington: United States Office of Education, 1938.

sample of 1660 state, county, and city courses of study revealed that only 13 per cent contained no statement of objectives. To be of maximum helpfulness in either teaching or testing, the objectives should be stated as specifically as possible. The expected pupil behavior must be indicated. It is not enough to say that the objective is "good citizenship" or "an integrated personality." These large indefinite terms must be broken down and stated in usable form.

With the list of teaching objectives for the course clearly and specifically stated, the teacher is ready to consider what procedures will be most appropriate for evaluating progress made toward the attainment of each objective. In other words, the teacher attempts to test what he has tried to teach by using techniques best adapted to each objective.

One writer[10] suggests that the objectives of instruction may be grouped into eight major categories:

1. Functional information
2. Various aspects of thinking
3. Attitudes
4. Interests, aims, purposes, appreciations
5. Study skills and work habits
6. Social adjustment and social sensitivity
7. Creativeness
8. A functioning social philosophy.

Another classification[11] recognizes ten major types:

1. The development of effective methods of thinking
2. The cultivation of useful work habits and study skills
3. The inculcation of social attitudes
4. The acquisition of a wide range of significant interests
5. The development of increased appreciation of music, art, literature, and other aesthetic experiences
6. The development of social sensitivity
7. The development of better personal-social adjustment
8. The acquisition of important information
9. The development of physical health
10. The development of a consistent philosophy of life.

For any given course these objectives must be expressed in terms of the specific changes in pupils which the teacher is seeking to bring about. A rather detailed inventory of the particular facts, principles, concepts and skills of the course is required, as well as the specific mental processes the pupil is expected to employ.

---

[10] Louis E. Raths, "Evaluating the Program of a School," *Educational Research Bulletin,* 17: 57–84, March 16, 1938.

[11] Smith, Tyler, and staff, *op. cit.,* page 18.

To measure whether these processes are really functioning, the teacher's inventory just mentioned must be presented to the pupils in language different from that of the text and class discussion, and opportunities must be offered to apply or to relate the objectives to new problems and situations. The center of gravity is the behavior of pupils rather than subject matter. The teacher must never confuse ends and means. The true relationship has been stated as follows: "The real ends of instruction are the *lasting* concepts, attitudes, skills, abilities and habits of thought, and the improved judgment or sense of values acquired; the detailed materials of instruction—the specific factual content—are to a large extent only a means toward these ends." [12]

A group of English teachers, for example, were able to recognize seven different aspects of "appreciation of literature." They then suggested the following ways [13] in which these aspects of appreciation may manifest themselves in pupil behavior:

1. *Satisfaction in the Thing Appreciated:* Appreciation manifests itself in a feeling, on the part of the individual, in keen satisfaction in, and enthusiasm for, the thing appreciated. The person who really appreciates a given piece of literature finds in it an immediate, persistent, and easily renewable enjoyment of extraordinary intensity.
2. *Desire for More of the Thing Appreciated:* Appreciation manifests itself in an active desire on the part of the individual for more of the thing appreciated. The person who really appreciates a given piece of literature is desirous of prolonging, extending, supplementing, renewing his first favorable response toward it.
3. *Desire to Know More about the Thing Appreciated:* Appreciation manifests itself in an active desire on the part of the individual to know more about the thing appreciated. The person who really appreciates a given piece of literature is desirous of understanding as fully as possible the significant meanings which it aims to express and of knowing something about the genesis, its history, its locale, its sociological background, its author, etc.
4. *Desire to Express One's Self Creatively:* Appreciation manifests itself in an active desire on the part of the individual to go beyond the thing appreciated, to give creative expression to ideas and feelings of his own which the thing appreciated has chiefly engendered. The person who really appreciates a given piece of literature is desirous of doing for himself, either in the same or in a different medium, something of what the author has done in the medium of literature.
5. *Identification of One's Self with the Thing Appreciated:* Appreciation manifests itself in the individual's active identification of himself with the thing

---

[12] E. F. Lindquist, "The Use of Tests in the Accreditation of Military Experience and in the Educational Placement of War Veterans," *Educational Record,* 25: 366, October, 1944.

[13] Louis Raths, "Appraising Certain Aspects of Student Achievement," *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I,* pages 114–115. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1938. This reference also contains some very suggestive tests designed to measure appreciation of literature, attitudes held toward important social issues, and some important aspects of thinking.

appreciated. The person who really appreciates a given piece of literature responds to it very much *as if* he were actually participating in the life situations which it represents.

6. *Desire to Clarify One's Own Thinking with Regard to the Life Problems Raised by the Thing Appreciated:* Appreciation manifests itself in a desire on the part of the individual to clarify his own thinking with regard to specific life problems raised by the thing appreciated. The person who really appreciates a given piece of literature is stimulated by it to rethink his own point of view toward certain of the life problems with which it deals and perhaps subsequently to modify his own practical behavior in meeting these problems.

7. *Desire to Evaluate the Thing Appreciated:* Appreciation manifests itself in a conscious effort on the part of the individual to evaluate the thing appreciated in terms of such standards of merit as he himself, at the moment, tends to subscribe to. The person who really appreciates a given piece of literature is desirous of discovering and describing for himself the particular values which it seems to hold for him.

Arnold[14] has shown that critical thinking can be taught in the elementary school and that various phases of the process can be measured. His study assumed that critical thinking involves the intelligent use of data, which was defined as the "ability to recognize relevance, dependability, bias in source, and adequacy of data in regard to a particular problem, question, or conclusion." The following item has to do with the recognition of bias in data:

Three boys were talking about whether or not a boy Jim was really "out" in a game of baseball that had been played that afternoon. John was on Jim's side. George was on the other side. Bill was not playing but was watching the game. Which of the three boys is most likely to be right? ______________.
Why? ____.

Recognition of the adequacy of data was measured by such test situations as the following:

Some people were talking about a ball team. Someone asked the others to tell why they thought this team was good. Here are the answers. Read them carefully. Put *B* before the one of these three that you think is the best reason for thinking this is a good ball team. Put *P* before the one you think is poorest, and put *F* before the one you think is just fair.

_____ 1. I think this is a good ball team.

_____ 2. I have seen them play once, and I think they are good players.

_____ 3. I have seen them play several times, and they are good players.

Read these next two. Find the better reason of the two; place a *B* before it. Find the poorer reason; place a *P* before it.

_____ 1. A man who studies baseball and writes much about it said they were a good team.

---

[14] Dwight L. Arnold, "Testing Ability to Use Data in the Fifth and Sixth Grades," *Educational Research Bulletin,* 17: 255–259, 278, December 7, 1938.

_____ 2. A man I talked to on the street the other day said they were good players.

Some boys were talking about a boy they called *D*. Jim said, "I saw *D* take a pencil from another pupil's desk. This makes me think he is a thief." If this is all that Jim knew about this, do you think Jim is right in thinking that *D* was a thief?

Put a line under your answer: YES NO AM NOT SURE

Now tell why you answered as you did.

____

____

It must be recognized, moreover, that some of the objectives of instruction cannot be measured by paper-and-pencil tests of any kind. At times rating scales, check lists, and other devices for recording observations of the individual at work or play are required. The term "test" in this discussion includes any instrument that affords valid evidence of progress made by pupils toward the attainment of the objectives of instruction.

One of the least tangible objectives of instruction is *creativeness.* Grimes and Bordin [15] have proposed that creative expression in art should result in the development of certain personality traits. These traits would be evaluated by the art teacher through observation during a conversation with his pupils. This process is guided by a check list upon which the teacher's record is entered. The writers suggest that this technique would be more valuable if a group of teachers co-operated in the construction of a check list of their own, rather than adopted wholesale the list given below:

I. *Initiative*—willingness to go into the unknown, to start off on a new track, to attempt something never attempted before; perseverance after recognizing a dead end; willingness to try again.

1. Attempts a medium, technique, or subject never attempted before.
2. Does not accept as final the view of the subject which he happens to have in the place where he begins, but moves around and views subject to be painted from many angles before deciding where to work.
3. Does not take for granted the posed object to be painted, but views the total situation and sees what, for him and his experience, there is in it that is paintable.
4. Assists in posing the model or arranging still life, and the like.
5. Brings material to be painted, as still-life objects, and the like, to the studio from outside sources.

---

[15] James W. Grimes and Edward Bordin, "A Proposed Technique for Certain Evaluations in Art," *Educational Research Bulletin,* 18: 1–5, 29, January 4, 1939.

6. Starts to work rather than depending on the teacher for instructions as to how to proceed.
7. Does not mind making mistakes but pursues the work despite reverses and difficulties.
   - *a*) Scrapes off in painting in oils the thick paint when canvas is gummed up; in water color washes out, when these steps are necessary for a continued work on the painting.
   - *b*) Does not regard initial drawing for a composition as absolute, but moves shapes as the developing experience demands adjustments and changes.
8. Participates in class discussions and contributes ideas and experiences.
9. Pursues some meaningful activity as sketching, if he finishes before the others in the class, rather than stalling around.
10. Does not demand approval or supervision from the teacher or other students at almost every step in his work.
11. Places his work away from himself, changes viewpoint in order to get a more objective view of his work.

II. *Concentration, Interest, Motivation*—vigor with which an individual attacks a problem and his oneness of purpose which would result in excluding factors nonrelevant to a given problem (perseverance is implied here).

1. Not distracted by others coming and going or talking.
2. Does not come and go himself.
3. Does not idly converse about matters exterior to the situation.
4. Does not let the work of others distract him from his own problems.
5. Does not quickly come to a dead end in his own work.
6. Does not stall around pretending to be working on a project.
7. Does work outside of the class that has bearing on class work, as: go to gallery, sketch, draw, and consult reproductive material.
8. Works on painting after hours.
9. Requests information as to work relative to his development.
10. Contributes to class discussion.
11. Talks to friends about his work and attempts to explain what he has been accomplishing and learning.
12. Paints the same subject more than once; does not give out quickly as regards subject matter.
13. Does not continually consult the time.

III. *Judgment*—weighing the factors in a situation and taking them into account before initiating a new action; that is, considering the possible results of an action before the initiation of it, seeing the social implications of a proposed action. Implied in this is knowing when to go off into the unknown and when not to; knowing when to pursue an independent course and when not to.

1. Does not aid others to the point of interfering with the progress of his own work or that of others.
2. Attempts to understand the point of view of others rather than thinking of ways to justify himself.
3. Selects a position to work which does not obstruct another's view of the subject.
4. Does not talk so loudly that it distracts others who are working.
5. Takes into consideration the desires and interests of others when arranging a subject to be painted.
6. Analyzes his working situation in relationship to time, when there is a group-determined time limit.
7. Knows when to pursue a project further and when to discard it and start another.
8. Uses materials and cares for them efficiently—cleans palettes, washes brushes, and the like.
9. Works plastically; that is, he allows for working with the forms rather than setting down an architectural plan as a rigid drawing which is filled in with color. Shows evidence of an exploratory and feeling-out attitude rather than a rigid method of working.
10. Examines critically criticism by others and makes use of it only so far as he feels it significant; that is, he reacts to it in terms of its validity rather than emotionally.
11. Takes into consideration the needs of others when using group materials.
12. Avoids vacillation in following out his own painting rather than shifting in style, execution, and attitude, as he sees others in the class going in a direction different from his own.

IV. *Co-operation*—the willingness to work in a group as a member of it in relationship to the teacher, individual members, and the whole group.

1. Makes use of criticism, does not react to it as personal insult, or cry, show anger, or leave class.
2. Is willing to alter his personal objectives to meet the situation.

It must be kept in mind that the objectives of the course represent *directions* of progress rather than *destinations* to be arrived at by individual pupils at any particular time. As far as possible, the progress of each individual should be measured in terms of his own interests, needs, and abilities. This is the aim of the modern school. The degree to which it is actually attained in any particular situation is dependent upon the resources available as well as upon the educational philosophy and skill of the teaching staff.

2. *The test should reflect the approximate proportion of emphasis in the course.* To insure a reasonable balance in the test, it is essential to draw up in outline form a sort of "job analysis" or "table

of specifications." This will guide the test maker much as the architect's blueprint and specifications guide the building contractor. It is well to indicate not only the various objectives the teacher has had in mind, but also, at least roughly, the relative amount of emphasis each objective has received in the actual teaching of the course. For example, the same test might not be equally valid for two teachers of a course in general science using the same textbook. This would be the case if one teacher emphasized almost altogether the memorization of isolated facts, while the other was much more concerned with the understanding of facts in relation to other facts, and in their application to practical problems in the community. The test should attempt to reflect faithfully the teaching emphasis. The amount of time devoted by the teacher to the various topical divisions of the course is a rough indication of what he considers to be their relative importance. The content of the test should show a similar proportion. The time devoted to a topic can at best indicate only the *number* of items to be included, and not the *type* of the items. The type of items to be used will depend upon the nature of the objective to be measured. A topical outline is only a partial guide to test construction. The table of specifications should also indicate the approximate teaching emphasis from the standpoint of knowledge, skills, attitudes, and other types of objectives that have been sought.

3. *The nature of the test must take into consideration the purpose it is to serve.* Any test is valid to the degree that it serves a specific purpose. If the purpose of the test is to afford a basis for school marks or for classification, it will attempt to rank the pupils in order of their total achievement. But if the purpose of the test is diagnosis, its value will depend upon its ability to reveal specific weaknesses in the achievement of individual pupils. Diagnostic tests would cover a limited scope but in much greater detail than a test of general achievement, and would be arranged so as to reveal the scores on the separate parts. The range of difficulty of the items and the discriminating value of the items individually are relatively less important in diagnostic tests. This is also true of mastery tests administered at the end of a teaching unit to determine when the minimum essentials have been achieved.

4. *The nature of the test must take into consideration the conditions under which it is to be administered.* In planning the test, attention must be given to such factors as the time available for testing, the facilities for duplicating the tests, and the cost of the materials, as well as the age and experience of the pupils being tested.

B. Preparing the Test

The second step is the actual preparation of the test. It has been found from experience that the following rules or suggestions are helpful:

1. *The preliminary draft of the test should be prepared as early as possible.* Many teachers find it desirable to jot down items to be included in the tests day by day as the teaching progresses. This is reasonable assurance that no important point in the course will be omitted in the test. If this is not done, the supplementary material of the course which is not included in the textbook and which may be of unusual value is especially likely to be overlooked. This practice also permits the material to "grow cold" and consequently to be more correctly appraised before it is included in the final draft of the test.

2. *As a rule, the test should include more than one type of item.* A variety of test types is likely to be more interesting to the pupil than a single form. This is especially true of long tests. Moreover, the requirement that the type of test situation should be the one which is most appropriate to the material to be included will usually necessitate that from two to four forms of objective items be used. These objective items are frequently combined with one or more discussion questions to make up the test.

3. *The content of the test should range from very easy to very difficult for the group being measured.* Ideally, the most reliable measurement for a given individual is afforded by a test made up of items of equal difficulty upon which he is capable of making 50 per cent of the maximum. In general achievement tests designed for measuring pupils of unequal ability, however, items should vary in difficulty from those which can be answered by almost 100 per cent to those which can be answered by just above 0 per cent of the pupils. This range is necessary in order to provide some items so easy that the weakest pupils can get 50 per cent of them correct and others so difficult that the strongest get only 50 per cent of them correct. For maximum discrimination the difficulty of the entire test should be such that, when allowance is made for chance, the average pupil in the group makes about 50 per cent of the possible score. It is clear, then, that a test which is of ideal difficulty for one class may be much too easy or much too difficult for other classes.

A few exceptions to this principle should be noted. In speed tests in such subjects as arithmetic and typewriting, where the objective is *rate* rather than *power,* all items should be of equal difficulty. The technique of making such tests is so complex that the ordinary teacher should rarely attempt them. Also, in both mastery and diagnostic tests the content is determined primarily by the *importance* of the subject matter rather than its *difficulty*. An adequate diagnostic test in the fundamental combinations in addition, for example, might yield average scores almost perfect in a strong class, and scores well below 50 per cent in a weak class.

4. *It is usually desirable to include more items in the preliminary draft of the test than will be needed in the final form.* This will permit "culling out," later on, items that may appear weak or not needed to produce the proper balance in the test. Ruch suggested that from 25 to 50 per cent more items be prepared than are likely to be required.[16]

5. *After some time has elapsed, the test should be subjected to a critical revision.* Then the items should be carefully checked with the table of specifications to see that the test shows the desired proportion of emphasis. A careful reading of the test after an interval of time will usually reveal some objectionable items. It is a good plan to have the test criticized by another teacher of the same subject. In this way some items are likely to be found which cover points of doubtful importance, others which are not clearly stated, and perhaps others about which there is disagreement as to the answers. The wording of the items should receive critical attention, particularly to avoid ambiguity. One serious error is the wording of items so that more than one reasonable interpretation is possible. The trouble with such ambiguous items is that a certain answer is correct with one interpretation, but with another interpretation a different answer is reasonably correct.

6. *The items should be so phrased that the content, rather than the form of the statement will determine the answer.* A common mistake is to include a telltale word or phrase that affords an unwarranted clue to the answer. These so-called *specific determiners* are especially common in true-false items.[17] It has been found that statements containing emphatic words, such as the adverbs "always," "never," "entirely," "absolutely," "exclusively," and the like, are much more likely to be false than true. On the other hand, words or expressions that limit the statement, such as "may," "sometimes," "as a rule," "in general," and the like, are much more likely to be true than false. Either these expressions should be avoided entirely, a suggestion which is rarely feasible, or

---

[16] G. M. Ruch, *The Objective or New-Type Examination,* page 154. Chicago: Scott, Foresman & Company, 1929.

[17] See I. H. Brinkmeier and G. M. Ruch, "Minor Studies in Objective Examination Methods: III—Specific Determiners in True-False Statements," *Journal of Educational Research,* 22: 110–118, September, 1930.

items containing them should be carefully balanced so that approximately the same number are true as false. Avoiding the language of the text will prevent pupils with good rote memories from answering items they may not understand. Sometimes clues are afforded by the spelling or by the grammatical form of the item. It is not unlikely that one of the reasons why many pupils prefer objective tests to other types is that such tests often contain items so worded as to be answered from a minimum knowledge of the subject matter involved. Such defects, however, are not inherent in objective testing; they can be avoided by the alert test maker. Administering the test to persons unfamiliar with the content of the course will often reveal those items which can be answered from general intelligence or from a general knowledge of language forms and usage.

The opposite mistake is often made also. Figurative language, needlessly heavy vocabulary, or involved sentence structure may so obscure the meaning of an item that it is marked incorrectly by pupils who really understand the point. Bob Burns' story of the time Grandpa Snazzy was a witness in court illustrates this error:

> The attorney says: "Now, Mr. Snazzy, did you or did you not, on the date in question or at any time previously or subsequently, say or even intimate to the defendant or anyone else, whether friend or mere acquaintance or in fact a total stranger, that the statement imputed to you, whether just or unjust and denied by the plaintiff, was a matter of no moment or otherwise? Answer—did you or did you not?"
>
> Grandpa thought a while and then says, "Did I or did I not what?"

Unless the test aims specifically to measure reading ability or general intelligence, the form of the item should neither impose unreasonable obstacles in the pupil's way nor provide clues which are too obvious. Both defeat the purpose for which the test was intended. A valid test item can be answered by an individual possessing a proper knowledge of the course and by nobody else.

7. *The items should be so worded that the whole content functions in determining the answer, rather than only a part of it.* There is often a wide discrepancy between what actually determines the pupil's response to a test and what the teacher intended. One of the principal reasons for this discrepancy is that only a part of the content of the item functions, the rest being wholly inert as far as the pupil is concerned. Lindquist [18] gives some excellent examples of this difficulty. Two of these, given below, should make the problem clear. Note the first:

---

[18] Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *The Construction and Use of Achievement Examinations,* pages 73–81. Boston: Houghton Mifflin Company, 1936.

The leader in the making of the compromise tariff of 1833 was (1) Clay, (2) Webster, (3) Jackson, (4) Taylor, (5) Harrison.

That the majority of the pupils who responded to this item correctly did so on the superficial basis of the strong verbal association between the words "compromise" and "Clay" is evidenced by the fact that fewer than half of them responded correctly when the item appeared in the following form:

The leader in the tariff revision of 1833 was (1) Clay, (2) Webster, (3) Jackson, (4) Taylor, (5) Harrison.

That the matching type of test is also subject to this error is shown by the next illustration:

Directions: Below are two columns of items. Match the items in the two columns by placing on the line before each group of words in Column A the right *number* from Column B.

Column A	Column B
.....1. a Phoenician contribution to civilization.	1. Mason and Dixon Line
.....2. most famous building of the ancient Greek world.	2. Spanish Armada
.....3. the fleet whose defeat in 1588 gave England the control of the Atlantic Ocean.	3. Saratoga
.....4. a boundary between two colonies that later became famous as the division between free and slave territory.	4. Dred Scott Decision
.....5. the victory which caused France to come to our aid during the Revolutionary War.	5. Parthenon
.....6. the law that forbade slavery north of the Ohio River.	6. Missouri Compromise
.....7. a ruling by the Supreme Court which opened all territory to slavery.	7. Alphabet
	8. Printing Press
	9. Ordinance of 1787

In most of the above items a single word gives the clue. For example, "boundary" in item 4 suggests "line" in response 1. Likewise, either "ruling" or "court" in item 7 suggests "decision" in response 4. If a pupil knows that "armada" means "fleet," he would be able to match item 3 with response 2 without knowing the date, the country, or the event involved. It should be noted, furthermore, that the above test would still be poor, even if each item were perfectly worded, because the items included are so diverse in character.

The test maker should attempt to anticipate the specific mental processes the pupil will employ in each response. For each item the teacher should raise such questions as the following: Are there any parts of the item that the pupil may disregard entirely and yet respond correctly? What is the minimum amount of knowledge required for a correct response?

8. *All the items of a particular type should be placed together in the test.* Sometimes completion, true-false, and multiple-choice items of varying numbers of choices are thrown together in random order. This arrangement is rarely, if ever, desirable. It is good practice to place together the items designed to measure a particular teaching objective. Such an arrangement not only facilitates the scoring of the test and the interpretation of the scores, but enables the pupil to take full advantage of the mind-set imposed by a particular test form.

9. *The items in the test should be arranged in order of difficulty.* It is especially important to have the easiest items at the beginning and the hardest ones at the end of the test. The exact order of the intervening items is less important. It will be recalled that one of the problems of measurement is to arrange conditions so that the thing being measured is disturbed as little as possible in the act of measuring. The psychological justification for placing the easiest items first is that such an arrangement has a wholesome effect upon the morale of the pupils taking the test. On the other hand, placing very difficult items at the beginning is likely to produce needless discouragement in the pupils, particularly with those of average ability and below. If the most difficult items come toward the end of the test, only the more capable pupils will probably get to them. After all, the only function of such items is to discriminate among the high-ranking pupils. In any event, any disturbing influence on the weaker pupils will come too late to affect seriously the results.

In advance of an actual tryout of the test, it is impossible to determine anything more than a rough estimate of the true difficulty order of the items, unless one is willing to go to the trouble of obtaining the pooled judgment of three or more persons. The judgment of a single experienced teacher regarding the difficulty of the items is likely to have considerable validity.[19] In any case it is usually possible to pick out those that will be at the extremes of the scale; and fortunately this is what is needed most. In later revisions of the test, the items can be placed in more exact order of difficulty.

---

[19] H. E. Smith, "The Validity of Teachers' Judgments of Difficulty in Curricular Materials," *Journal of Educational Psychology,* 21: 460–466, September, 1930.

10. *A regular sequence in the pattern of responses should be avoided.* The order of responses should be a chance order rather than a regular pattern. If items are arranged alternately true and false, or two true and two false, for example, the pupil is likely to discover the arrangement. To facilitate scoring, it is sometimes suggested that multiple-choice items be so arranged that the correct responses give combinations easy to remember, such as a familiar date like 1453. But there is always risk that the pupil will "get the hang" of the pattern and answer successfully without considering the content of the item at all.

11. *Provision should be made for a convenient written record of the pupil's responses.* Such a record is a check list, a rating scale, or some other similar form upon which the observer makes a systematic and permanent record of a pupil's behavior under a given set of conditions. It is particularly difficult to provide a satisfactory written record of responses on the oral quizzes. In the ordinary test the pupil makes his own record in writing either on the test paper or a specially prepared answer sheet. The problem then is merely that of arranging the test so that the labor of scoring will be reduced to a minimum. Such devices as numbering the responses in multiple-choice items and the blanks in completion items, so that the responses will be recorded in a column rather than scattered irregularly over the page, save time and reduce the chances of error in scoring. Merely grouping the items by fives, rather than spacing them uniformly, reduces somewhat the eyestrain in scoring the test.

12. *The directions to the pupil should be as clear, complete, and concise as possible.* The aim should be to make the instructions so clear that the weakest pupil in the group knows what he is expected to do, although he may not be able to do it. The pupil should be told how and where to mark the items, the time allowed to do so, and any reduction for errors to be made in scoring. The amount of detail required will depend upon the maturity of the pupils and their experience with that particular type of test. To very young children, for example, it will be better to say "draw a line under" rather than "underline," and "draw a ring around the right answer" rather than "encircle the correct response." In the lower grades it is usually desirable for the teacher to read the directions aloud to the pupils while they follow silently the written directions on their test papers. Wherever the form of the test is unfamiliar or complicated, a generous use of samples correctly marked, and fore-exercises or practice tests that do not count in determining the score is to be recommended. Sometimes a blackboard demonstration is the best way to make the procedure clear. As the pupils become

familiar with the various types of items and the procedure used in scoring them, the directions may be greatly abridged.

A single illustration should make these points clear. The following directions may be considered reasonably satisfactory for a class unfamiliar with the objective tests:

Directions to the Pupil: Below are thirty statements about measurement in education. Examine each statement and decide whether it is true or false. In the ( ) before each statement you think is true, put +; in the ( ) before each statement you think is false, put 0. You will have ten minutes for the test. Your score will be the number right minus the number wrong. Study the samples below. They are answered correctly.

SAMPLES:
(0) A. High reliability insures high validity in a test.
(+) B. Group tests of intelligence originated in America.

After the pupils have become familiar with true-false tests and the method employed in scoring them, the directions may be shortened to a form somewhat as follows:

Directions: In the ( ) before each item put + if true, and 0 if false. You will have ten minutes for the test.

One other point warrants consideration. Should pupils be told or encouraged to guess at items about whose answers they are in doubt? Some authorities would require the pupils to attempt all items on recognition tests. They would include some such statement as, "If you do not know, guess!" Others would go to the other extreme and say, "Do not guess!" Still others, perhaps the majority, would be content with informing the pupil that the correction formula [20] is to be employed and let him use his judgment about attempting doubtful items. Unfortunately, the experimental evidence on this point is neither extensive nor altogether convincing. Most of the studies have merely attempted to compare the relative effect of the first two practices upon the validity and reliability of the scores, without considering the third possibility at all.

The results have usually favored the do-not-guess instructions by a slight margin.[21] However, Votaw,[22] Lentz,[23] Cronbach,[24] and others have found some evidence of a factor of "acquiescence" operating in taking tests since do-not-guess instructions placed ascendant students at an advantage on recognition tests over submissive students. It is argued that such instructions reduce the validity of achievement tests, since they become in some degree measures of personality traits. Where the correction formula is used, this difficulty is apparently not very serious. Some investigators have also found that good students tend to improve their scores when they attempt doubtful items, whereas poor students do not. Regardless of whether or not the pupil is instructed to guess, the use of the correction formula tends to improve the validity of recognition tests, although the gain is hardly sufficient to justify the labor where the number of responses involved for each item is four or more.

---

[20] This formula is discussed on page 122.

[21] For a brief summary of the experimental literature on this problem, see: Walter W. Cook, "Achievement Tests," in *Encyclopedia of Educational Research,* edited by Walter S. Monroe, pp. 1292, 1298. New York: The Macmillan Company, 1941.

[22] David F. Votaw, "The Effect of Do-not-guess Directions upon the Validity of True-false or Multiple-choice Tests," *Journal of Educational Psychology,* 27: 698–703, December, 1936.

[23] Theodore F. Lentz, "Acquiescence as a Factor in the Measurement of Personality," *Psychological Bulletin,* 35: 659, November, 1938.

[24] Lee J. Cronbach, "Studies of Acquiescence as a Factor in the True-false Test," *Journal of Educational Psychology,* 33: 401–415, September, 1942.

All things considered, the author offers the following recommendations:

a. The use of recognition tests with fewer than four responses to each item should be avoided wherever possible.

b. Regardless of the number of possible responses, the score should probably be the number right on all tests used with pupils below the junior-high-school level.[25]

c. When tests with only two or three possible responses to each item are used with pupils above the sixth grade, the correction formula should be employed.

d. Whenever the correction formula is to be used, the pupils should be so informed.

e. The theory of the correction formula and the experimental evidence regarding the handling of doubtful items should be discussed with pupils before the test begins. They should then be allowed to use their best judgment, without being specifically advised by the directions to guess or not to guess.

C. Trying Out the Test

After the test has been prepared according to plan, it is ready to be given a trial in actual use. Since it is impossible in advance to know exactly how good the test is or to locate all the poor items, the tryout should be considered a necessary step in constructing the test in its final form. The following four principles should govern the tryout. With the possible exception of the second, these principles are equally applicable to the later use of the test in its final form.

---

[25] Admittedly this decision is based on practical considerations, rather than on experimental evidence. It is usually easier to make the test a little longer than to explain to young children the logic of the formula, and they are likely to be suspicious of what they do not understand.

1. *Every reasonable precaution should be taken to insure normal conditions for the test.* This is important because the responses to any test are partly determined by the conditions under which it is given, as well as by the test itself. It is usually well to have the test administered to the pupils in the familiar environment of their own classroom. Any tendency to cheat should be forestalled by careful supervision. Where cheating is likely to be a special problem, pupils may be so seated that every other seat is vacant, or the test items may be arranged in different orders for pupils seated close together.

2. *The time allowance for the test should be generous.* This is more important in the tryout than in the later use of the test in its final form. One reason for this is that the items are arranged at best in only a rough order of difficulty, and, if the time allowance is too short, pupils may not have time to try items toward the end of the test, which they may be capable of answering correctly. Short time allowances should be avoided, therefore, in order to secure the data needed for determining the difficulty and the discriminating value of the items. What time allowance is to be considered generous will depend upon the purpose of the test and upon the ability and experience of the pupils. For example, it is obvious that the time limits of speed tests should be so short that even the best pupil does not have time to finish the test. On the other hand, more time should be allowed for diagnostic tests than for tests of general achievement; and tests of a purely factual character can be answered more quickly than those involving the higher mental processes.

Lindquist suggests that, in general achievement tests, the time allowance should be so adjusted that "at least 75 per cent of the pupils will have time at least to *consider* all items in each section." [26] Ruch seemed to favor time limits "so that 90 per cent can attempt all items within their power." [27] In accordance with this standard, Ruch suggested that for fairly short items of a factual character, three recall or four recognition items per minute is a "reasonable expectancy for upper-elementary and high-school pupils." For reasoning tests the corresponding time allotments would be increased for recall items to one or two items per minute, and for recognition items to two or three items per minute. Younger pupils and longer items would demand still more time.

---

[26] Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *op. cit.*, page 116.

[27] G. M. Ruch, *op. cit.*, page 312.

The above standards have in mind the requirements for the ordinary use of the test in its final form, rather than for the tryout, for which more time should be allowed. Since so many factors influence the time demands of a particular test, the writer suggests that in the tryout sufficient time be allowed so that all, or practically all, the pupils have time to finish. If the examiner will record during the progress of the test the percentage of the pupils who are still at work after various amounts of time have elapsed, the information will be useful in determining the time allowances for later revisions of the test.

3. *The scoring procedure adopted should be as simple as possible.* As a rule, the best procedure in scoring objective tests is to give one point of credit for each correct response. In recognition tests this means one point for each item properly marked, and in recall tests it means one point for each blank correctly filled. It is unnecessary to weight the items according to estimated difficulty or importance. Even in essay examinations weighting is much less important than is ordinarily assumed. Practically all pupils will be in the same rank order regardless of the weighting of the individual items.[28]

The use of the correction formula is probably justified where chance alone would make possible guessing the correct response to one half or one third of the items. The general formula is usually written:

$$S = R - \frac{W}{n-1}.$$

In this formula

S is the score corrected for guessing.
R is the number of right responses.
W is the number of wrong responses.
n is the number of responses presented for each item.

For two-response tests this becomes

$$S = R - W.$$

For three-response tests the formula is

$$S = R - \tfrac{1}{2}W.$$

Similar formulas can be derived for recognition tests with a larger number of possible responses, but the increase in validity and reliability rarely justifies the labor.

---

[28] *Cf.* Alexander J. Phillips, "Further Evidence Regarding Weighted Versus Unweighted Scoring of Examinations," *Educational and Psychological Measurement,* 3 151–155, Summer, 1943.

4. *Before the actual scoring begins, answer keys and scoring rules should be prepared.* In teacher-made objective tests satisfactory scoring keys can be prepared by simply filling in the correct responses, preferably with a colored pencil, on one of the unused tests. Scoring then consists of comparing the pupil's responses with those on the key placed beside his paper. In essay examinations the key consists of a model paper containing a complete set of answers, together with the points to be allowed on each. Definite rules are necessary to secure uniformity in scoring. The rules for scoring objective tests usually say merely that one point will be allowed for each correct response and that no fractional credits will be allowed, and indicate whether or not the correction formula will be used. The rules for essay examinations give the weight for each question, and tell whether or not any deductions are made for errors in spelling, language usage, and so forth. In mathematics tests the rules should cover such points as whether or not the answers must be reduced to lowest terms, whether or not credit will be allowed for solutions correct in principle but with the wrong answer, and the like.

D. Evaluating the Test

After the papers have been scored, the results should be interpreted and evaluated from two points of view: first, as to the quality of the test itself; and, second, as to the quality of the pupils' responses. While the ultimate interest of the test maker is in the light thrown by the test results upon the quality of the teaching and organization that exists in the school, his first concern should be the quality of the test used. It should be apparent that only tests of high merit afford information of value regarding the school situation. To what extent, then, does the test possess the three characteristics of satisfactory measuring instruments, validity, reliability, and usability? Only the last of these can be confidently determined in advance. The five principles that follow are suggested for evaluating the test from the viewpoint of its validity and reliability.[29] If the test is found to possess these qualities in high degree, the scores should then be carefully analyzed for their value in instruction and school administration. If the test is found to lack these qualities, the scores can be disregarded and the test subjected to a thorough revision. No matter how carefully the test is prepared in the first place, its merits should be established and not merely assumed.

---

[29] For a brief but suggestive report, see: Ellis Weitzman and Walter J. McNamara, "Techniques Used in Analyzing the Learning Achievement of Naval Aviation Cadets," *Journal of Educational Psychology,* 35: 181–185, March, 1944.

1. *The difficulty of the test is a rough indication of its validity.* The difficulty of the test as a whole is determined by finding what percentage the average score made is of the maximum possible score. In general achievement tests, the nearer this average is to 50 per cent, the better. The difficulty of the individual test items is obtained by finding the percentage of successful responses for each item. Items answered by 100 per cent or by 0 per cent of the pupils are of no value in a test of general achievement. The difficulty of the test is relatively unimportant in mastery tests and in diagnostic tests.

2. *The validity of the individual items in the test is determined by their ability to discriminate between pupils who rank high and those who rank low on the test as a whole.* There are several methods of determining the validity of test items. Only the simplest of these methods are practical for use with informal tests. A satisfactory procedure for the classroom teacher is to determine the percentage of correct responses (or of incorrect responses) to each test item by the pupils who rank in the highest fourth of the class on the test as a whole, and to compare with the corresponding percentage in the lowest fourth of the class.[30] The items in which the percentage of correct responses of the high group exceeds that of the low group by the largest amount are best; those in which percentages are the same are useless; and those in which the percentage of correct responses of the high group falls behind that of the low group are detrimental. Items showing zero or negative discrimination should be either reworded or thrown out altogether.

3. *It is a good practice to have the items interpreted or criticized by persons who have taken the test.* It is impossible to anticipate fully all the mental processes pupils will employ in responding to a test item. These can be determined only by making inquiry of pupils who have taken the test. In this way irrelevancies and ambiguities will be revealed that were wholly unsuspected by the maker of the test. Often a slight change in wording is sufficient to remedy the difficulty. At other times the item must be entirely discarded. If a test contains too many of these items, the scores on the test should not be counted in determining the pupil's record in the class. Inviting members of the class to assist in this critical evaluation of the test helps to create a favorable attitude toward the measurement process employed by the instructor, and is a valuable educational experience in itself.

---

[30] For a defense of the selection of the contrasting groups from the 25 to 27 per cent at the extremes of the distribution, see: Truman L. Kelley, "The Selection of Upper and Lower Groups for the Validation of Test Items." *Journal of Educational Psychology,* 30: 17–24, January, 1939.

4. *Whenever possible, the results on the test should be checked against an outside criterion.* For short tests covering small units of subject matter, this process is likely to be difficult and of little value. Even here it is sometimes helpful to compare the ranks of the pupils on the test with those assigned by the teacher before the test is given. The validity of the longer and more important tests can be determined in a more satisfactory manner by comparing the scores of the pupils on each test with their scores on a good standard test covering the same material and given at about the same time. The coefficient of correlation [31] obtained between the two series of scores is the most exact method of expressing the amount of agreement, although a rough indication can be obtained by comparing the percentage of scores which lie in the same fourths of the two series of scores.

5. *It is sometimes desirable to obtain the reliability coefficient of the test.* The author recognizes that it is possible to overestimate the value of the reliability coefficient. The makers of standardized tests have often made this mistake. However, the reliability coefficient does have some merit in evaluating informal tests, although the value is mainly negative. Low reliability coefficients indicate tests of doubtful merit, but high reliability coefficients *per se* do not establish the value of the tests. To be of real value these coefficients must be supported by other criteria.

Since informal tests are likely to have but one form, the best method of obtaining the reliability coefficient, which is a measure of the internal consistency of the test, is by correlating the scores on the even-numbered items with those on the odd-numbered items. This is the reliability of the half test. The reliability of the whole test is then estimated by the use of the Spearman-Brown formula.[32]

The construction of an informal teacher-made test, then, involves these four steps: planning, preparing, trying out, and evaluating. It is perhaps more correct to say that these activities constitute a cycle in the construction of a test, for it is often necessary to repeat these steps, particularly the last three, several times before the test is brought to its finished form.

SELECTED REFERENCES FOR FURTHER READING

Broom, M. E., *Educational Measurements in the Elementary School.* New York: McGraw-Hill Book Company, Inc., 1939. Chapter VI.

Greene, Harry A., Jorgensen, Albert N., and Gerberich, J. Raymond, *Measurement and Evaluation in the Secondary School.* New York: Longmans, Green & Company, 1943. Chapter VIII.

---

[31] See Chapter VIII, pages 244–245.

[32] *Ibid.*

Hawkes, Herbert E., Lindquist, E. F., and Mann, C. R., *The Construction and Use of Achievement Examinations.* Boston: Houghton Mifflin Company, 1936. Part I.

Lee, J. Murray, *A Guide to Measurement in Secondary Schools.* New York: D. Appleton-Century Company, 1936. Chapter X.

Odell, C. W., *Traditional Examinations and New-Type Tests.* New York: D. Appleton-Century Company, 1928. Chapters III, VII, and IX.

Orleans, Jacob S., *Measurement in Education.* New York: Thomas Nelson and Sons, 1937. Appendix D.

Rinsland, Henry Daniel, *Constructing Tests and Grading in Elementary and High School Subjects.* New York: Prentice-Hall, Inc., 1938. Chapter IX.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman & Company, 1929. Chapters VII, XI, and XII.

Smith, Eugene R., Tyler, Ralph W., and Evaluation Staff, *Appraising and Recording Student Progress.* New York: Harper & Brothers, 1942. 550 pages.

Tyler, Ralph W., *Constructing Achievement Tests.* Columbus: Bureau of Educational Research, Ohio State University, 1934. 110 pages.

CHAPTER V

Principles of Constructing Specific Types of Objective Tests

A. Introduction

Types of objective tests. The principal types of objective test items used by classroom teachers may be listed as follows:

1. Recall types.
    a. Simple-recall.
    b. Completion.

2. Recognition types.
    a. More common.
        (1) Alternative-response.
        (2) Multiple-choice.
        (3) Matching.

    b. Less common.
        (1) Rearrangement.
        (2) Identification.
        (3) Analogy.
        (4) Incorrect statement.

This chapter will consider the uses and limitations of the commonly used forms of objective tests and suggest rules which have been found to be of value in constructing them. It will also give illustrative items in a variety of fields, drawn mainly from standard tests.

Frequency of use by teachers. Two studies present data on the frequency of use by classroom teachers of various forms of test items. In the first of these studies Conneau[1] analyzed 45,418 test items that appeared in 375 objective examinations submitted in a prize contest. This study doubtless represented the practice of superior teachers in 1928, rather than that of average teachers. In 1936 Lee and Segel[2] reported an analysis of the types of informal tests used by 1,600 high-school teachers distributed widely over the United States. That there is rather surprising agreement between these studies is indicated by Table 6. In both studies the completion form ranks first and the true-false second. Conneau grouped all recall forms under completion, while Lee and Segel separated out the one-word-answer type. This type of item, which ranked third in the latter study, has not been included in the table. The next most popular item is the multiple-choice form. The most striking disagreement is in the relative rank of the essay examination. In the earlier study only .6 per cent of the questions were of the essay type, while in the more recent study 16 per cent of the teachers appear to be using that type extensively. This apparent revival of interest in the essay examination is probably less marked than the difference in ranks between the two studies would indicate, since the earlier tests were written for prize competition. In fact, Lee and Segel[3] conclude that there has been a definite shift toward objective tests.

---

[1] Summarized by G. M. Ruch, *The Objective or New-Type Examination,* pages 188–190. Chicago: Scott, Foresman & Company, 1929.

[2] J. Murray Lee and David Segel, *Testing Practices of High School Teachers,* pages 6–12. United States Office of Education Bulletin, No. 9, 1936.

Table 6

RANKINGS OF TEST ITEMS ACCORDING TO FREQUENCY OF USE AS REVEALED BY TWO STUDIES

Type of Item	Conneau	Lee and Segel
Completion	1	1
True-false	2	2
Multiple-choice	3	4[a]
Essay	11	5
Problems	7	6
Matching	4	7

[a] The one-word answer form was included under completion by Conneau, but was recognized as a type by Lee and Segel. In the latter study it ranked third.

Comparative validity and reliability of various types of tests. Ruch[4] summarized the experimental studies available in 1929 and came to the conclusion that "the new-type tests are at least as valid as the essay examinations," and that the various objective types are "not greatly unequal in validity." Ruch also concluded that "for equal working times recall and recognition types are not greatly dissimilar," although recall tests tended to rank at the top and true-false at the bottom in most of the studies.

During the next ten years several experimental studies and excellent summaries of the literature were published. Those by Kinney and Eurich[5] and by Lee and Symonds[6] are the most comprehensive. The latter study points out that the problem of determination of the comparative merits of different measuring instruments is not only "one of the most important" it is also "one of the most poorly done."

---

[3] J. Murray Lee and David Segel, *op. cit.*, page 6.

[4] G. M. Ruch, *op. cit.*, pages 281–306.

Rinsland[7] also summarized the experimental literature to 1938 and suggested two cautious conclusions:

> 1. One might conclude that the objective tests, with probably the exception of the true-false type, are as valid as, or perhaps slightly more valid than, the essay or subjective examination; and that, of all the objective forms, the completion or simple recall seems to be the most valid.
>
> 2. Generally speaking, the various types of objective tests have about equal reliability when compared on the basis of working time. Differences of reliability may be due primarily to the wording of individual items rather than to the objective form.

Lindquist[8] takes the position that many of the studies which have attempted to determine the comparative validities and reliabilities of various test forms have been "inconclusive, if not definitely misleading." He points out that these comparative studies have not always recognized the specific nature of test validity, have overemphasized the importance of reliability, and have often failed to control such factors as relative skill in constructing the various test forms and the time allotments for the tests. In view of these limitations Lindquist comes to the conclusion that "in making a selection from a number of test techniques in any specific test situation or in relation to any specific objective of instruction, the test constructor must, at present, depend almost entirely upon logical considerations rather than upon the experimental or empirical evidence that is now available."

A summary published in 1941[9] concludes that "few dependable generalizations can be drawn from studies in this area" and that the "over-lappings are much more significant than minor differences between averages." Another article in the same volume[10] arrives at this conclusion:

---

[5] L. B. Kinney and A. C. Eurich, "A Summary of Investigations Comparing Different Types of Tests," *School and Society,* 36: 540–544, October 22, 1932.

[6] J. Murray Lee and Percival M. Symonds, "New Type of Objective Tests: A Summary of Recent Investigations," *Journal of Educational Psychology,* 24: 21–39, February, 1933; 25: 161–184, March, 1934.

[7] Henry Daniel Rinsland, *Constructing Tests and Grading in Elementary and High School Subjects,* pages 295–299. New York: Prentice-Hall, Inc., 1938.

[8] Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *The Construction and Use of Achievement Examinations,* pages 97–103. Boston: Houghton Mifflin Company, 1936.

[9] Max D. Engelhart, "Examinations," in the *Encyclopedia of Educational Research,* edited by Walter S. Monroe, pages 471–476. New York: The Macmillan Company, 1941.

> The relative effectiveness of a test technique is specific rather than general. . . . It is probable that the validity of a test technique in most fields is more a function of the ingenuity with which it is applied than it is of the test technique employed. . . .
>
> Adequate comparisons between test techniques can therefore be made only for specific material when results are used for a specific purpose, when items are constructed with specific insight and ability, and on the basis of validity coefficients computed for equal amounts of testing time when each test is administered at its optimum rate.

To be of practical guidance to the classroom teacher, research should seek answers to such specific questions as the following: In the measurement of what specific objectives in science is the true-false technique of most worth? What testing technique is most effective for measuring vocabulary in a foreign language? What distinctive value, if any, has the rearrangement test in history? It seems too bad that, for the present, one's choice of the tools of science must depend upon one's personal judgment and general educational philosophy rather than upon direct experimental evidence.

It is well to recognize that knowledge may exist and function on at least four different levels. The lowest level involves mere *recognition.* A person's general reading vocabulary, as distinguished from his speaking and writing vocabulary, is an example of knowledge where the ability to recognize is the important thing. The next higher level involves *recall.* For knowledge of many types to have value, one must be able to recall it when needed. Familiar examples are one's speaking and writing vocabulary, the names and faces of acquaintances, and the ordinary number combinations in arithmetic. Sometimes one needs to recall separate facts or isolated bits of knowledge, but at other times the organization is important. The person who is an entertaining conversationalist, an interesting letter writer, or an effective public speaker must be able to present his knowledge in a connected form. A still higher level of knowledge involves the ability to *interpret and evaluate.* At this level the learner must have a sufficient understanding of the material to be able to see it in its relationships to other things. The exercise of discrimination and judgment is implied. The highest level of all involves *application.* The person who is able to utilize information acquired in one situation and who applies it to the intelligent solution of problems in a new setting has arrived at true mastery.

---

[10] Walter W. Cook, "Achievement Tests," in the *Encyclopedia of Educational Research,* pages 1290–1291.

It seems reasonable to assume that the type of test used must be appropriate to the level of knowledge being measured.[11] Recognition tests of the multiple-choice and matching types appear adequate for the first level of knowledge. Recall tests would seem to be required for the other three levels. Wherever relationships and organization are important, the essay type is more appropriate than the simple recall. However far more important than the *type* of test is the skill with which it is used. Understanding, evaluation, and many aspects of thinking can be measured by recognition tests, but to do so requires a degree of skill that the regular classroom teacher rarely attains.[12] It is also probably true that most recall tests measure memory only.

B. Simple-Recall Tests

Definition. The *simple-recall* test is here somewhat arbitrarily defined as one in which each item appears as a direct question, a stimulus word or phrase, or a specific direction. The response must be *recalled* by the pupil from his past experience rather than merely *identified* from a list of suggested answers supplied by the teacher. The simple-recall test is differentiated from the essay examination primarily upon the basis of length of response required; the typical response to the simple-recall item is short, preferably a single word or phrase.

Advantages and limitations. This type of test has the obvious advantage of familiarity and "naturalness." It also stimulates desirable study practices and almost completely eliminates guessing as a factor for measurement, thus avoiding two of the most common faults of objective tests. The *simple-recall* test is particularly valuable in mathematics and the physical sciences, where the stimulus appears in the form of a problem requiring computation.[13] It also has wider application to test situations presented in the form of maps, charts, and diagrams in which the pupil is required to supply, in spaces provided, the names of parts keyed by numbers or letters.

One limitation of the simple recall test is that it tends to measure highly factual knowledge, consisting of isolated bits of information. Also the scoring is somewhat laborious and not always entirely objective. These limitations need not be very serious when the tests are carefully prepared as can be seen from the illustrations which follow.

---

[11] For a stimulating discussion of this point, see: Douglas E. Scates, "Complexity of Test Items as a Factor in the Validity of Measurement," *Journal of Educational Research,* 30: 77–92, October, 1936.

[12] For a comprehensive discussion of this problem, see: William A. Brownell and Committee, "The Measurement of Understanding," *Forty-Fifth Yearbook of the National Society for the Study of Education, Part I.* 338 pages. Chicago: University of Chicago Press, 1946.

[13] Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *op. cit.,* pages 363–367.

I. Illustrations of Simple-Recall Tests

Below are a few sample test items of the simple-recall form that have been taken from well-known standard tests.[14] Excellent examples of this and other test forms used in a variety of school subjects on all educational levels are to be found in Rinsland.[15]

New Stone Reasoning Tests in Arithmetic[16]

1. James had 5 cents. He earned 13 cents more and then bought a top for 10 cents. How much money did he have left? *Answer:* ____________
2. How many oranges can I buy for 35 cents when oranges cost 7 cents each? *Answer:* ____________

Sones-Harry High School Achievement Test, Part II[17]

1. What instrument was designed to draw a circle? ...... (____________)1
2. Write "25% of" as "a decimal times." .... (____________)2
3. Write in figures: one thousand seven and four hundredths (____________)3

Cooperative General Mathematics Tests for College Students, Form 1934[18]

28. How many axes of symmetry does an equilateral triangle have? .... (   )
29. Eight is what per cent of 64? .... (   )
30. Write an expression that exceeds M by X. .... (   )
31. Solve the formula $V = \frac{Bh}{3}$ for h. .... (   )

---

[14] In the examples of the various types of objective tests that follow an effort has been made to illustrate a wide variety of mechanical arrangements of items as well as of subject matter. It is recognized that they are not all of equal merit.

[15] Henry Daniel Rinsland, *op. cit.*, pages 23–222.

[16] Devised by C. W. Stone, and published by Bureau of Publications, Teachers College, Columbia University.

[17] Devised by W. W. D. Sones and David P. Harry, Jr., and published by World Book Company.

[18] Devised by H. T. Lundholm and L. P. Siceloff, and published by Cooperative Test Service.

Iowa Placement Examinations, Chemistry-Training[19]

1. The atomic weight of K is 39; of Cl, 35.5; of O, 16.
   What is the molecular weight of $KClO_3$? ....
2. If 7 gm. of iron unite with 4 gm. of sulphur, how many gm.
   of iron sulphide will be produced? ....

Tests on Everyday Problems in Science, Unit XII[20]

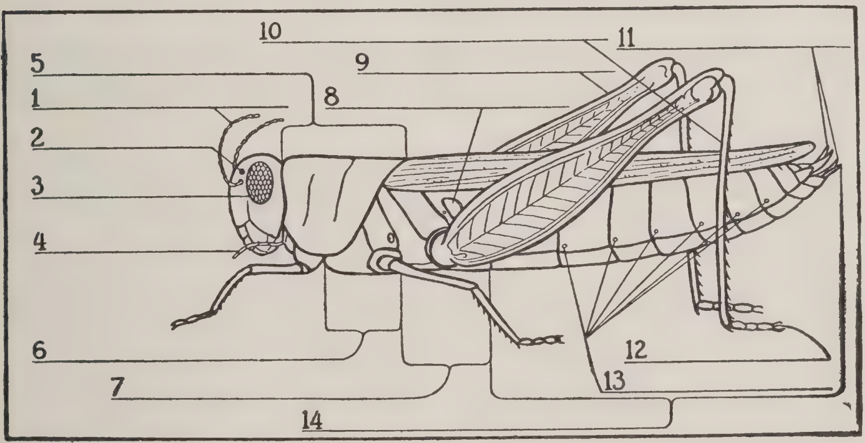
What device is used in a vacuum-cleaner to pump air into the dust bag? .... (15)____________

What is the pressure in pounds of ordinary air per square inch? .... (16)____________

What does the word "pneumatic" mean? .... (17)____________

An Exercise from a Biology Workbook[21]

DIRECTIONS: As you locate each part using a hand lens on an actual specimen, find the corresponding part in the accompanying illustration and label it. Consider how each part functions in the life of the grasshopper.



Parts of the Grasshopper

The following items from an informal class test illustrate the possibilities of recall tests with more than one response to each item:

---

[19] Devised by G. D. Stoddard and J. Cornog, and published by Extension Division, State University of Iowa.

[20] Devised by C. J. Pieper and W. L. Beauchamp, and published by Scott, Foresman & Company.

[21] Prepared by Arthur O. Baker and Lewis H. Mills to accompany their *Dynamic Biology Today*. Chicago: Rand McNally & Company, 1943.

For each event below give the country, date, and person with whom you associate it:

Event	*Country*	*Date*	*Person*
First psychological laboratory......	___________	________	___________
First general intelligence test......	___________	________	___________
First standardized achievement test	___________	________	___________

II. Rules and Suggestions for Construction

The simple-recall is one of the most familiar test forms and one of the easiest to prepare. The main problem is how to phrase the test situations so that they will call forth responses of a higher intellectual level than mere rote memory, and so that they can be scored with a minimum expenditure of time and effort.

1. *The direct-question form is usually preferable to the statement form.* It is more natural for the pupil and is likely to be easier to phrase.

EXAMPLE: The first president of the United States was ____________
BETTER: Who was the first president of the United States? ____________

2. *The questions should be so worded that the response required is as brief as possible, preferably a single word, number, symbol, or at most a short phrase.* This will objectify and facilitate scoring.

3. *The blanks provided for the responses should be in a column, preferably at the right of the questions.* This arrangement facilitates scoring and is more convenient for the pupil. The illustrations above show various ways of arranging the answer column.

4. *The use of textbook language in wording the question should be reduced to the minimum.* Unfamiliar phrasing will reduce the possibility of correct responses that represent mere meaningless verbal associations, and also will eliminate the temptation of pupils to memorize the exact language of the book.

5. *The questions should be so worded that there is only one correct response.* This is a standard which is difficult to reach, since pupils are marvelously resourceful in reading into questions interpretations which the teacher never intended. For example, the question in ancient history, "Name two ancient sports" elicited this reply from an ingenious student, "Antony and Cleopatra." This possibility would not have arisen had the question taken this form: "What were two popular athletic contests in ancient Greece?" All acceptable replies which are based on any legitimate interpretation of the question should receive credit, and must be listed on the scoring key. A little extra care in wording the question will save much time and trouble later.

C. Completion Test

Definition. The *completion* test may be defined as a series of sentences in which certain important words or phrases have been omitted and blanks submitted for the pupil to fill in. A sentence may contain a single blank, or it may contain two or more blanks. The sentences in the test may be disconnected, or they may be organized into a paragraph. Each blank counts one point.

Advantages and limitations. The mental processes which the pupil must employ in supplying the responses required in completion tests are very similar to those required in simple recall tests, although perhaps on a somewhat higher level. It is not surprising that the advantages and limitations of these two types of tests are also similar. The completion test has wide applicability, as far as subject-matter is concerned, but unless prepared with extreme care is likely to measure rote memory rather than real understanding; or it may turn out to be more a measure of general intelligence or linguistic aptitude than of school achievement.

The scoring is likely to be even more laborious than that of simple-recall tests. This is not only because the scoring is somewhat subjective, but also because the missing words are written in blanks scattered all over the page, rather than in a column. While these limitations cannot be entirely eliminated they can be greatly reduced, as is evident from the illustrations below.

I. Illustrations of Completion Tests

Stanford Achievement Test, Paragraph Meaning, 1940 Edition[22]

Directions: [Abridged] Write JUST ONE WORD on each line. *Be sure to write each answer on the line that has the same number as the missing word in the paragraph.*

1–2–3 Answer
In olden days men made their own pens from the quills
of feathers. It required considerable skill to cut a pen properly
so as to suit one's individual taste in writing. Students were 1___________
always on the lookout for good goose, swan, turkey, or other bird
feathers. Goose quills made the most satisfactory —1— for 2___________
general —2—, but schoolmasters liked pens made from the —3—
of swan feathers because they fitted best behind the ear. 3___________

Public School Attainment Tests for High School Entrance[23]

3. *Question:* Did this team have a coach?

Answer: No, they taught (3) how to play without any coach. (3)............

4. *Q:* Did all of you have matches?

A: Of course! Each one had (4) own water-proof box full. (4)............

---

[22] Devised by Truman L. Kelley, Giles M. Ruch, and Lewis M. Terman, and published by World Book Company.

[23] Devised by Henry D. Rinsland and Roland L. Beck, and published by Public School Publishing Company. (The form of completion with all responses in a column, instead of staggered within sentences, was devised by Rinsland.)

Tests of Everyday Problems in Science, Unit XI[24]

A pry-pole is an example of a machine called the.......... (11).............
A capstan is an example of a machine called the.......... (12).............
A screw is an example of a machine called the.......... (13).............
Your teeth are examples of machines called .... (14).............

Gregory Tests in American History[25]

	Write your words and dates here.
2. The man who headed the first expedition to circumnavigate the globe was	2 ....
7. The Articles of Confederation were in force from 1781 to	7 ....
9. The "Old Liberty Bell" rang out the decision of congress to be free from England in the year ...	9 ....

Cooperative English Test, Form 1934, Series 1[26]

20. Write on the lines to the right the contractions—shortened forms to represent how the words are naturally spoken—for the seven groups of words underlined in the following sentences. For instance, for *do not,* you would write *don't.* You need not copy the sentence, but only the seven contractions.

I have read his story, but I *cannot* ______________________
believe that *he will* get a passing ______________________
grade on it, for *it is* not well written — ___________________
and *has not* a clear-cut plot. The char- ______________________
acters *are not* at all interesting; *they* ______________________
are not even human. ______________________

II. Rules and Suggestions for Construction

Most of the suggestions made for constructing simple-recall tests apply equally well to completion tests. The dangers to be avoided are largely the same for both forms. A few suggestions may be offered, however, that have special reference to completion items. The main problems of constructing completion tests are three in number: (1) How to phrase the statements so as to indicate the *type* of response desired; (2) How to avoid giving the pupil unwarranted clues to the specific responses expected; and (3) How to arrange the items so as to facilitate scoring. The first two suggestions below apply to problem one, the next five apply to problem two, and the last five suggestions are for problem three. In short, a good completion statement gives a *reasonable basis* for determining the response desired without providing *unwarranted clues,* and is arranged to *facilitate scoring.*

---

[24] Devised by C. J. Pieper and W. L. Beauchamp, and published by Scott, Foresman & Company.

[25] Devised by C. A. Gregory, and published by C. A. Gregory Company.

[26] Devised by Sterling A. Leonard and others, and published by Cooperative Test Service.

1. *Avoid indefinite statements.* The pupil is entitled to know the *type* of response desired, and when this is done the scoring is far more rapid.

EXAMPLE: Abraham Lincoln was born in ____________.
BETTER: Abraham Lincoln was born in the state of ____________.
The date of Abraham Lincoln's birth was ____________.

The first statement fails to indicate whether the desired response is the date, the place, or the circumstances of his birth. In that form, legitimate answers might be "February" or "1809," for the date; "Kentucky," or possibly "The South," for the place; and "poverty," or "a log cabin," for the circumstances of his birth. By a slight change in wording the statement is made quite definite.

2. *Avoid overmutilated statements.* If too many key words are left out, it is impossible to know what meaning was intended.

EXAMPLE: The ____________ is obtained by dividing the ____________ by the ____________.

In its present form, it is impossible to tell whether the statement refers to educational measurement or to arithmetic. And, if the former, it might refer to IQ, EQ, or AQ.

BETTER:

1. The EQ is obtained by dividing the ____________ by the ____________.
2. The ____________ is obtained by dividing the ____________ by the MA.

3. *Omit key words and phrases, rather than trivial details.* If this is not done the response may be as obvious as the first example below, or as unnecessarily difficult as the second example.

EXAMPLES:

1. Abraham Lincoln was born February ____________, 1809.
2. Abraham Lincoln was born in ____________ County, Kentucky.

4. *Avoid lifting statements directly from the text.* This puts too great a premium upon rote memory.

5. *Whenever possible, avoid "a" or "an" immediately before a blank.* These words unnecessarily limit the responses that can be used in the blank.

EXAMPLE: Mary picked an ____________ off the tree and ate it.

BETTER: Mary ate the ____________ which she picked off the tree.

It is apparent that the words "pear," "peach," "plum," "cherry," "lemon," "pineapple," and the like could not be used in the first statement. In fact, the choice tends to narrow down to two familiar fruits, "apple," or "orange." Moreover the second statement contains no specific determiner.

6. *Make the blanks of uniform length.* If the blanks vary in length the pupil has a clue to the length of answer expected. Even more of a clue is afforded by using a dot or a dash for each letter in the correct word.

EXAMPLE:

1. The second president of the United States was ____ _____ from the state of _____________.
2. The president in office during the Mexican War was _____ _ ____ from the state of _________.

BETTER:

1. The second president of the United States was ____________ from the state of ____________.
2. The president in office during the Mexican War was ____________ from the state of ____________.

7. *Avoid grammatical clues to the answer expected.*

EXAMPLE: The authors of the first performance test of intelligence were ____________.

BETTER: The first performance test of intelligence was prepared by ____________.

8. *Choose statements in which there is only one correct response for the blanks.* The scoring is far more objective if only one specific word or phrase can be used to complete the statement.

9. *The required response should be a single word or a brief phrase.* The more the scorer has to read the more time will be required.

10. *Arrange the test so that the answers are in a column at the right of the sentences.* The illustrations above show various ways in which this may be done. When each sentence contains but a single blank, the scoring is made easier if the blank comes at the end. The Tests of Everyday Problems in Science and the Gregory Tests in American History are examples. If the sentences contain more than one blank, the scoring is more rapid if the blanks are numbered and the pupil is directed to write his responses in the correspondingly numbered blank in the answer column at the right. Rinsland[27] suggests that the following wording of the directions will be clear to all pupils above the fourth grade, although it may be necessary to explain the word "correspondingly" in grades four to seven.

Directions: In each of the sentences below, one or more words, numbers, or dates are needed in the numbered blank spaces to make the sentences complete and true. Place the word or words in the correspondingly numbered blank to the right.

11. *Prepare a scoring key which contains all acceptable answers.* Although it is desirable to have only one response which can be considered correct for each blank, this is not possible in all cases. As a rule, a satisfactory key can be made by writing in red the correct answers on a copy of the test.

12. *Allow one point for each blank correctly filled.* Avoid fractional credits and unequal weighting of items on the basis of difficulty or importance.

---

[27] Henry Daniel Rinsland, *op. cit.*, page 56.

D. Alternative-Response Tests

Definition. An *alternative-response* test is made up of items each of which admits of only two possible responses. The usual form is the familiar true-false test. Other common forms are right-wrong, correct-incorrect, yes-no, and same-opposite.

Advantages and limitations. Obvious advantages of the alternative-response test are its apparent ease of construction, applicability to a wide range of subject-matter, complete objectivity of scoring, and its wide sampling of knowledge tested per unit of working time. The true-false test, a form very popular with classroom teachers, has been the object of more research and of more criticism than any other form of objective test. The negative-suggestion effect and the factor of guessing are often pointed out as limitations of this type of test. While the use of the correction formula appears to make a satisfactory adjustment in the total score, the alternative-response is not well adapted to educational diagnosis. The danger of negative-suggestion when pupils see statements which are false has apparently been overestimated, but perhaps it is wise not to use true-false tests as pretests or with young children. In such cases it is better to avoid the alternative-response test, or to use a question that can be answered by *yes* or *no* instead of with a declarative statement.

Several modifications and alleged improvements of the true-false test have been proposed. Barton,[28] for example, has suggested crossing out the part of the statement that is in error, while other studies[29] have shown that having pupils correct the wrong statements increases the reliability of the test.[30] Still others[31] have proposed that items be weighted according to the judgment of the pupil, or be marked *true, false, doubtful.* All of these suggestions add somewhat to the labor of scoring and have not received wide acceptance. Furthermore, strictly speaking, when these modifications are followed, the test is no longer of the alternative-response type. As a rule, the most obvious way to "improve" the true-false test is also the best; that is, *make the test longer and prepare it more carefully.* At least 75 items are desirable, and 50 may be set as an absolute minimum, unless the test covers a very narrow range or is used for instructional purposes only. One advantage of the true-false test is that it can cover more items in the same time than any other test type.

---

[28] W. A. Barton, Jr., "Improving the True-False Examination," *School and Society,* 34: 544–546, October 17, 1931.

[29] Ernest E. Bayless and Ralph C. Bedell, "A Study of Comparative Validity as Shown by a Group of Objective Tests," *Journal of Educational Research,* 23: 8–16, January, 1931; F. D. Curtis, W. C. Darling, and N. H. Sherman, "A Study of the Relative Values of Two Modifications of the True-False Test," *Journal of Educational Research,* 36: 517–527, March, 1943; W. H. E. Wright, "The Modified True-False Item Applied to Testing in Chemistry," *School Science and Mathematics,* 44: 637–639, October, 1944.

[30] This type of modified true-false test is illustrated in the workbook.

[31] Kate Hevner, "A Method for Correcting for Guessing in True-False Tests and Empirical Evidence in Support of It," *Journal of Social Psychology,* 3: 359–362, August, 1932.

Should pupils be advised to look over true-false tests and change the answers on doubtful items? Several studies have attempted to answer this question. Hill[32] made an extensive investigation of the problem and came to the conclusion that there is "not much advantage to be gained by changing one's answers on a true-false test," although the advantage was somewhat greater in changing from true to false than in the reverse. There is some evidence that the better pupils profit most from rechecking and revising their work. Even if the scores are not always improved, it is probably a good work habit to encourage.

The low esteem in which test experts hold the alternative-response type of test, especially the true-false form, is indicated by the infrequency with which it has appeared in recent standardized achievement tests. This is due chiefly to its weakness as an instrument of diagnosis, and to the fact that such tests must be made much longer than other objective tests in order to secure comparable reliability. Although this type of test has been overworked by classroom teachers, it does have a legitimate, even if restricted, use in informal tests. For example, the true-false test seems well adapted to testing the persistence of popular misconceptions and superstitions. Ordinary alternative-test situations are encountered in which it is difficult or impossible to make enough plausible responses for a multiple-choice test. There are many troublesome situations of this sort in language usage. Common examples include the case forms in pronouns, correct use of singular and plural verbs, confusions of past tense and past participles, the use of *sit* and *set, lay* and *lie,* and many others. A safe rule would be to restrict the use of the alternative-response test to those situations to which other test forms are inapplicable, and then to give particular care to the wording of the items.

---

[32] George E. Hill, "The Effect of Changed Responses in True-False Tests," *Journal of Educational Psychology,* 28: 308–310, April, 1937.

I. Illustrations of Alternative-Response Tests

Progressive Achievement Tests—Primary Battery[33]

DIRECTIONS: If the two words are the same or mean the same, write *S*. If they mean different things, write *D*.

6. saw ______________________ was
17. REPEAT ___________ RESEAT
20. presidential ________ *presidential*

Iowa Silent Reading Tests, New Edition, Sentence Meaning, Elementary[34]

DIRECTIONS: Read each question. If the answer is "Yes," fill in the space under YES in the margin. If the answer is "No," fill in the space under NO. Do not guess.

1. Is a dime less in value than a nickel? ....1 YES NO
2. Can you see things clearly in a fog? ....2 YES NO
3. Is geography studied in public schools? ....3 YES NO

Michigan Botany Test[35]

1. Is the corolla made up of petals? YES NO
2. Do all individual flowers contain both stamens and pistil? YES NO
3. Is the anther a part of the pistil? YES NO

Tests in English Fundamentals: Grammar[36]

DIRECTIONS: Classify the italicized words in the sentences below as **adjectives** or **adverbs** by placing check marks in the proper columns:

		Adjective	Adverb
3. That was a *silly* remark.	3		
6. Those flowers smell *sweet.*	6		
11. You can *hardly* expect him to wait.	11		

---

[33] Devised by Ernest W. Tiegs and William W. Clark, and published by California Test Bureau.

[34] Devised by H. A. Greene, A. N. Jorgenson, and V. H. Kelley, and published by World Book Company.

[35] Devised by O. W. Laidlaw and Clifford Woody, and published by Public School Publishing Company.

[36] Devised by R. Davis, and published by Ginn and Company.

The 1939 Iowa Every-Pupil Tests in Basic Skills[37]

Directions: In each of the following sentences there are two or more numbered words or phrases inclosed in brackets. If you think the *first* word or phrase is correct, place an *X* in the *first* box of the corresponding row on the answer sheet. If you think the *second* answer is correct, place an *X* in the *second* box of the proper row, etc.

7. Ted is $\left\{\begin{matrix}\text{1. a}\\ \text{2. an}\end{matrix}\right\}$ industrious man.

54. My father $\left\{\begin{matrix}\text{1. has}\\ \text{2. hasn't}\end{matrix}\right\}$ no money.

62. I want everyone to help $\left\{\begin{matrix}\text{1. himself}\\ \text{2. themselves}\end{matrix}\right\}$.

Cooperative Plane Geometry Test, Revised Series Q[38]

Directions: Read these statements and mark each one in the parentheses at the right with a plus sign (+) if you think it is always true, or with a zero (0) if you think it is always or sometimes false.

1. The opposite angles of a parallelogram are equal ..... 1( )

2. A diameter of a circle divides the circle into two equal parts 2( )

17. If two triangles are similar, their areas are in the same ratio as the medians drawn to corresponding sides .... 17( )

18. All similar polygons are equilateral .... 18( )

Tests on Everyday Problems in Science: Unit III[39]

Directions: There are 25 incomplete statements in this test, each followed by parts (*a*), (*b*), (*c*), and (*d*). One or more of these parts, or perhaps none of them, correctly complete the incomplete statement. You are to place a plus sign (+) in the parentheses (near the right margin) opposite each part which correctly completes the statement, and a minus sign (−) opposite each part which does not correctly complete the statement.

13. Minerals in our food supply
    (a) furnish heat and energy to the body .... ( )
    (b) are the only materials of which cells can be built .... ( )
    (c) are good regulators of certain of the body activities .... ( )
    (d) help particularly to build bone and blood .... ( )

---

[37] Devised by H. A. Greene, and published by Extension Division, State University of Iowa.

[38] Devised by Emma Spanney and L. P. Siceloff, and published by the Cooperative Test Service.

[39] Devised by C. J. Pieper and W. L. Beauchamp, and published by Scott, Foresman & Company.

Cooperative Solid Geometry Tests, Form 1934[40]

DIRECTIONS: Read these statements and mark each one in the parentheses at the right with a plus sign (+) if you think it is true, or with a zero (0) if you think it is false, wholly or in part.

4. Any number of planes may be passed through a given straight line ( )
27. Two planes parallel to the same straight line are parallel to each other ( )
41. The square of a diagonal of a cube is three times the square of its edge ( )

George Washington University English Literature Test[41]

T F 1. "Il Penseroso" describes the charms of a merry social life.
T F 4. "Pilgrim's Progress" is one of the greatest prose allegories in literature.
T F 8. In his poem "The Bells," Poe describes the process of making bells.

II. Rules and Suggestions for Construction

The true-false test is often thought to be one of the easiest types to prepare. This superiority is more apparent than real, however. Experienced test makers are convinced that no test form demands greater skill. Unusual care must be exercised in wording true-false statements so that the *content* rather than the *form* of the statement will determine the response. The aim should be to phrase the statement so as not to make its meaning needlessly obscure on the one hand, nor to provide unwarranted clues on the other. This balance requires a delicate skill of adjustment that is rare among makers of informal tests. The following specific suggestions may be found helpful in constructing true-false tests. Many of the suggestions for constructing multiple-choice tests that are found in the next section are also applicable here.

1. *Avoid specific determiners.* It has been found that strongly worded statements are much more likely to be false than true, while moderately worded statements are much more likely to be true than false. Examples of the former are those containing "all," "always," "never," "no," "none," "nothing," and the like; examples of the latter are those containing "may," "some," "sometimes," "often," "as a rule," and the like. If care is taken to balance the proportion of true and false items containing any particular expression, that expression ceases to be a specific determiner that affords a clue to the answer.

2. *Avoid a disproportionate number of either true or false statements.* Since several studies have shown that false statements are more valid than true statements, the suggestion is sometimes made that the test should have more false statements than true. If this were generally done, however, the validity of the

---

[40] Devised by H. T. Lundholm and others, and published by Cooperative Test Service.

[41] Devised by K. T. Omwake and others, and published by Center for Psychological Service.

false statements would probably be reduced, since the pupil would then tend to mark all doubtful statements false.

3. *Avoid the exact language of the textbook.* Lifting true statements directly from the textbook, or making false statements by changing a single word or expression puts too great a premium on rote memory.

4. *Avoid trick statements.* These are usually statements which appear to be true but which are really false because of some inconspicuous word or phrase.

EXAMPLES: 1. "The Raven" was written by Edgar Allen Poe.
2. The battle of Hastings was fought in 1066 B.C.
BETTER: 1. "The Raven" was written by Edgar Allan Poe.
2. The battle of Hastings was fought in 55 B.C.

Statements which are partly true and partly false should be avoided for the same reason.

5. *Avoid double negatives.* Such statements are especially bad, since pupils well versed in English grammar would recognize that two negatives equal an affirmative, while other pupils would interpret such statements as emphatic negatives.

6. *Avoid ambiguous statements.* With one interpretation the statement may be true and with another equally plausible interpretation it may be false. It is impossible to tell what is being measured when a statement has more than one legitimate interpretation.

7. *Avoid unfamiliar, figurative, or literary language.* The experience of the learner must be considered. A statement is badly worded when a pupil who understands the point involved misses it because of the language employed.

8. *Avoid long statements, especially those involving complex sentence structure.* Same reason as for the preceding suggestion.

9. *Avoid qualitative language wherever possible.* Quantitative language conveys more exactly the meaning intended. Expressions such as "few," "many," "large," "small," "old," "young," "important," "unimportant," are vague and indefinite.

EXAMPLES: 1. Shakespeare lived a long time ago.
2. Jamestown was settled a few years before Plymouth.
BETTER: 1. Shakespeare lived in the sixteenth century.
2. Jamestown was settled about twenty-five years before Plymouth.

10. *Require the simplest possible method of indicating the response.* Instead of requiring the pupil to write *True* and *False* or *Yes* and *No,* let him write *T* and *F, Y* and *N,* or underline the correct response. The symbols "+" for true and "0" for false are so distinct as to make scoring still easier. When the pupil must choose between two words or expressions, the responses should be numbered so that they can be indicated by writing the correct number.

11. *Indicate by a short line or by (   ) where the response is to be recorded.* The responses may be arranged in a column at either the left or right of the statements. Most authorities prefer the answers at the right.

12. *Arrange the statements in groups.* There is some advantage in scoring if the items are arranged in groups of five, with double spacing between each group.

E. Multiple-Choice Tests

Definition. A *multiple-choice* test is made up of items each of which presents three or more responses, only one of which is *correct* or definitely *better* than the others.[42] Each item may be in the form of a direct question, an incomplete statement, or a word or phrase. This form of test is to be distinguished from the *multiple-response* type, which requires that two or more responses be made to a single item.

Possibilities and limitations. The multiple-choice type of item is usually regarded as the most valuable and most generally applicable of all test forms. Lee regards it as "one of the best means for testing judgment that is available."[43] Lindquist asserts that it is "definitely superior to other types" for measuring such educational objectives as "inferential reasoning, reasoned understanding, or sound judgment and discrimination on the part of the pupil."[44]

One study[45] suggests fourteen types of questions which may be asked in multiple-choice test items. The list is not all-inclusive and does not intend to prescribe the exact language to be used but serves as a guide in formulating the questions.

1. Definition
   a. What means the same as . . . ?
   b. What conclusion can be drawn from . . . ?
   c. Which of the following statements expresses this concept in different form?

2. Purpose
   a. What purpose is served by . . . ?
   b. What principle is exemplified by . . . ?
   c. Why is this done . . . ?
   d. What is the most important reason for . . . ?

3. Cause
   a. What is the cause of . . . ?
   b. Under which of the following conditions is this true . . . ?

4. Effect
   a. What is the effect of . . . ?
   b. If this is done, what will happen?
   c. Which of the following should be done (to achieve a given purpose)?

---

[42] It is also possible, especially in English usage and spelling tests, to have several correct forms and only one incorrect or least desirable form, which is to be chosen in each item.

[43] J. Murray Lee, *A Guide to Measurement in Secondary Schools,* page 379. New York: D. Appleton-Century Company, 1936.

[44] Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *op. cit.*, page 138.

[45] Charles I. Mosier, M. Claire Myers, and Helen G. Price, "Suggestions for the Construction of Multiple-Choice Test Items," *Educational and Psychological Measurement,* **5**: 261–271, Autumn, 1945.

5. Association
   What tends to occur in connection (temporal, causal, or concomitant association) with . . . ?
6. Recognition of Error
   Which of the following constitutes an error (with respect to a given situation)?
7. Identification of Error
   a. What kind of error is this?
   b. What is the name of this error?
   c. What recognized principle is violated?
8. Evaluation
   What is the best evaluation of . . . (for a given purpose) and for what reason?
9. Difference
   What is the important difference between . . . ?
10. Similarity
    What is the important similarity between . . . ?
11. Arrangement
    In the proper order, (to achieve a given purpose or to follow a given rule) which of the following comes first (or last, or follows a given item)?
12. Incomplete Arrangement
    In the proper order, which of the following should be inserted here to complete the series?
13. Common Principle
    All of the following items except one are related by a common principle:
    a. What is the principle?
    b. Which item does not belong?
    c. Which of the following items should be substituted?
14. Controversial Subjects
    Although not everyone agrees on the desirability of ____________, those who support its desirability do so primarily for the reason that __________.

Unusual care must be exercised in the construction of multiple-choice tests in order to avoid the inclusion of irrelevant or superficial clues, and to insure that the tests measure something more than the memory of factual knowledge. The value of multiple-choice tests in diagnosis depends upon the skillful selection of the incorrect choices presented in the items.[46]

I. Illustrations of Multiple-Choice Tests

The items below, taken from well-known standard tests, illustrate several different arrangements of multiple-choice tests in a variety of subjects.[47] This type of test is widely used in all school subjects, on all educational levels, and for measuring a variety of teaching objectives.

---

[46] Ellis Weitzman and Walter J. McNamara, "Apt Use of the Inept Choice in Multiple-Choice Testing," *Journal of Educational Research*, 39: 517–522, March, 1946.

[47] These tests are not all equally good, however. The reader will note that some of them are not wholly consistent with the principles set forth in this chapter.

Special attention should perhaps be called to two of the illustrations, both of which are suggestive to teachers in making informal tests. The Nelson High-School English Test illustrates the possibility of testing punctuation with a minimum of scoring labor. The Cooperative Test of Social Studies Abilities, designed by Wrightstone, shows how objective tests may be used to test more than the memory for factual knowledge. This is a good example of a test of the pupil's ability to interpret facts—an ability which is an important aspect of thinking.

Columbia Research Bureau English Test[48]

Directions: Select the *one* correct spelling in each line, underline it, and put its number in the parentheses at the right.

2. 1 deceive 2 diceave 3 decieve 4 deceave .... ( )
27. 1 acurate 2 acurrate 3 accurate 4 accurrate .... ( )
37. 1 eroneous 2 erroneous 3 erronious 4 erroneus .... ( )

The Modern School Achievement Tests, Language Usage[49]

Directions: In each sentence, choose the word or group of words that make the best sentence. Then on the dotted line at the right, copy the number that is before the correct form.

4. I borrowed a pen 1. off 2. off of 3. from my brother. ....

7. Every student must do 1. your 2. his 3. their best. ....

17. He 1. has got 2. has 3. has gotten his violin with him. ....

The Barrett-Ryan Literature Test: Silas Marner[50]

A. ( ) An episode that advances the plot is the—1. murdering of a man. 2. kidnapping of a child. 3. stealing of money. 4. fighting of a duel.
B. ( ) Dolly Winthrop is—1. an ambitious society woman. 2. a frivolous girl. 3. a haughty lady. 4. a kind, helpful neighbor.
C. ( ) A chief characteristic of the novel is—1. humorous passages. 2. portrayal of character. 3. historical facts. 4. fairy element.

---

[48] Devised by H. R. Steeves, Allan Abbott, and Ben D. Wood, and published by World Book Company, 1926.

[49] Devised by A. I. Gates and others, and published by Bureau of Publications, Teachers College, Columbia University.

[50] Devised by E. R. Barrett, T. M. Ryan, and H. E. Schrammel, and published by Kansas State Teachers College, Emporia.

Wesley Test in Political Terms[51]

1. An embargo is
   1. a law or regulation 2. a kind of boat 3. an explorer 4. a foolish adventure 5. an embankment ( )
2. An injunction is a
   1. part of speech 2. wreck 3. union of two things 4. court order 5. form of advice ( )

Unit Scales of Attainment in Foods and Household Management[52]

2. The spoon should be placed
   1. at the top of the plate
   2. at the left of the fork
   3. in the spoon holder on the table
   4. at the right of the knife .... ( )

11. The best breakfast for a three-year-old child is
   1. prune pulp, oatmeal, milk, toast 2. prune pulp, oatmeal, coffee, toast 3. candy, oatmeal, milk, toast 4. toast, orange, cocoa, oatmeal .... ( )

40. We get the most calories per pound from
   1. proteins 2. carbohydrates
   3. fats 4. mineral matter
   5. vitamins .... ( )

Traxler Silent Reading Test, Word Meaning[53]

8. The *commendation* is deserved.
   (1) success (2) blow (3) popularity (4) good fortune
   (5) praise ( )
9. His actions received *condemnation.*
   (1) approval (2) applause (3) censure (4) sympathy
   (5) contempt ( )

Cooperative French Test, Junior Form 1936[54]

2. Quand on vous pose une question, il faut
   1 répondre, 2 se taire, 3 se sauver, 4 tourner le dos, 5 baisser la tête .... ( )
7. Cette dame est ma grand'mère; je suis
   1 son fils, 2 son neveu, 3 son frère, 4 son cousin, 5 son petit-fils .. ( )

16. J'ai deux frères, Jean et Paul. Jean a sept ans, Paul en a treize et moi j'ai douze ans. Qui est le plus jeune?
   1 Jean, 2 Paul, 3 moi, 4 dix ans, 5 les deux frères .... ( )

---

[51] Devised by E. B. Wesley, and published by Charles Scribner's Sons.

[52] Devised by Ethel B. Reeve and Clara M. Brown, and published by Educational Test Bureau, Inc.

[53] Devised by Arthur E. Traxler, and published by Public School Publishing Company.

[54] Devised by Jacob Greenberg and Geraldine Spaulding, and published by Cooperative Test Service.

Nelson High School English Test[55]

Directions: Some of the sentences contain errors in punction; some of them are correct. If you think some mark is not needed, cross out the letter indicating that mark under the word "Omit." If you think some additional mark is needed, cross out the letter indicating that mark under the word "Add." If you think the exercise is correct, cross out the letter r. Key: a—apostrophe; c—comma; d—dash; e—exclamation point; h—hyphen; p—period; q—quotation mark; s—semicolon.

	Add	Omit	Right
1. You must elect a chairman, three judges and an official timekeeper.	X h d s	q	r
6. He said "that either you or I must go."	c s d e	X	r
8. The car which John is driving is a new one.	d q s c	d	X
14. "Well, I think highly of them Mary" I said.	e h s X	p	r

Cooperative Test of Social Studies Abilities, Experimental Form Q[56]

INTERPRETING FACTS

Directions: The exercises in this part consist of a series of paragraphs each followed by several statements about the paragraph. In the parentheses after each statement, put a

1, if the statement is a reasonable interpretation, fully supported by the facts given in the paragraph;
2, if the statement goes beyond and cannot be proved by the facts given in the paragraph;
3, if the statement contradicts the facts given in the paragraph.

[The sample exercise and the explanation are omitted.]

I. The nineteenth century witnessed a rapid growth in Germany's industrial power. Like England, Germany came to have a fairly satisfactory balance between the amount of its export and import trade. Heavy exports of coke supplied full cargoes for ships to foreign ports and helped to balance heavy importations of raw materials. The imports especially provided a means for distributing freight rates to the advantage of the German trader competing overseas. By these means Germany was constantly obtaining larger portions of world trade. German wares were carried into every trading realm, and trade meant political as well as commercial power in foreign lands.

1. Through growth in foreign trade, Germany's industrial power increased in the nineteenth century . . . . 1( )
2. Germany had an export trade equal in volume to that of England 2( )
3. Germany exported very little coke to foreign countries . . . . 3( )
4. England was unable to balance the tonnage of her import and export shipments . . . . 4( )

---

[55] Devised by M. J. Nelson, and published by Houghton Mifflin Company.

[56] Devised by J. Wayne Wrightstone, and published by Cooperative Test Service.

5. By reducing freight rates Germany was constantly gaining a greater percentage of world trade .... 5( )
6. The sale of German wares in every part of the world resulted in added political influence and commercial growth .... 6( )

II. Rules and Suggestions for Construction

The purpose of suggestions 1 to 5 below is to avoid unwarranted clues to the desired response, the purpose of suggestions 5 to 10 is to encourage responses on a high intellectual level, and the purpose of suggestions 11 to 15 is to make the scoring as simple and rapid as possible.

1. *Make all optional responses grammatically consistent.* For example, if the verb is singular, avoid plural responses, and vice versa. Avoid using "a" or "an" as the word in an incomplete statement immediately preceding the list of responses, unless all options begin with consonant sound (in the case of "a") or all begin with a vowel sound (in the case of "an").

2. *As a rule, use direct questions rather than incomplete statements.* The question form is more natural and less likely to contain irrelevant clues.

3. *Avoid making the correct response consistently longer or shorter than the others.*

4. *Avoid using in the correct response the same words or phrases that occur in the question or incomplete statement.*

5. *Arrange the responses so that the correct one occurs in random order.* The pupils are likely to detect any regularly recurring pattern in the sequence of responses. When the number of correct responses in the various positions is kept even, there is some evidence [57] that the next-to-the-last choice tends to be slightly more difficult than the others.

6. *Make all responses plausible.* The aim should be to make each suggested response so plausible as to tempt pupils who have only superficial knowledge of the point involved. The plausibility of incorrect responses may be increased by using familiar, stereotyped, or textbook phraseology, or expressions very similar to those in the question or incomplete statement.

7. *At least four choices should be presented whenever possible.* Increasing the number of plausible choices tends to reduce the guessing factor. Horst [58] has found, however, that when the incorrect responses are of equal difficulty the chance element is less than when the choice is among a greater number of responses with a wider range of difficulty.

8. *In phrasing multiple-choice test items, consideration should be given to the fact that the answer may be arrived at by eliminating the incorrect responses as well as by selecting the correct response directly.* Requiring the pupil to select the *least* satisfactory response in the series given, or the one that is *not* true, will often compel a careful comparison of all the possible responses.

9. *In testing for the understanding of a term or concept, the term should usually be presented first, followed by a series of definitions or descriptions from which the choice is to be made.* If the order is reversed, so that from a series of

---

[57] Ellis Weitzman and Walter J. McNamara, "The Effect of Choice Placement on the Difficulty of Multiple-Choice Questions," *Journal of Educational Psychology,* 36: 103–113, February, 1945.

[58] Paul Horst, "The Difficulty of a Multiple-Choice Test Item," *Journal of Educational Psychology,* 24: 229–232, March, 1933.

terms the choice is made of the one that best fits the definition or descriptive statement, the selection frequently can be made based upon superficial verbal associations and not upon genuine understanding.

10. *To measure the higher levels of understanding, increase the homogeneity of the options provided.* The following illustration from Lindquist[59] shows how the degree of required discrimination increases with the homogeneity of the responses presented:

A. Engel's law deals with
  1. the coinage of money
  2. the inevitableness of socialism
  3. diminishing returns
  4. marginal utility
  5. family expenditures

B. Engel's law deals with family expenditures for
  1. luxuries
  2. food
  3. clothing
  4. rest
  5. necessaries

C. According to Engel's law, family expenditures for food
  1. increase in accordance with the size of the family
  2. decrease as income increases
  3. require a smaller percentage of an increasing income
  4. rise in proportion to income
  5. vary with the tastes of families

To respond correctly to *A*, all that is required is the knowledge that Engel's law deals with family expenditures. In *B* a knowledge of the specific item of expenditure is necessary. The maximum degree of discrimination, however, is required in *C*, where still more information is given.

11. *Require the simplest possible method of indicating a response.* This usually means that the responses are numbered and the choice is made by indicating the *number* of the response. In the first two or three grades where key numbers may not be understood, it will be better to permit the more natural response of *underlining* the correct answer.

12. *Indicate by a short line or by ( ) where the response is to be recorded.* The response column may be either at the left or right.

13. *Arrange the items in groups.* As a rule, groups of five will be suitable, although other numbers of items may sometimes be better. Double space between each group.

14. *Use the correction formula only if the number of choices is fewer than four.* If there are four or more responses suggested for each item, the gain in validity is seldom sufficient to warrant the labor of making corrections for chance.

15. *Group together all items with the same number of choices.* This is always desirable, and is imperative when the correction formula is to be used.

F. Matching Tests

Definition. A *matching* test typically consists of two columns, each item in the first column to be paired with a word or phrase in the second column upon some basis suggested. In the simplest form of matching test the number of responses is exactly the same as the number of items. Frequently, so-called "imperfect" matching tests are made which provide more responses than are required. Sometimes the items in the first column are incomplete sentences, each of which requires a word or phrase from the second column for its completion. Occasionally two, or even more, columns of responses are given, from each of which a choice must be made for each item in the first column. The matching test is also useful for identifying numbered places or parts on maps, charts, and diagrams.

---

[59] Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *op. cit.*, pages 146–147.

Advantages and limitations. There are many types of learning which involve the association of two things in the mind of the learner. Common examples are the following: Events and dates, events and persons, events and places, terms and definitions, foreign words and English equivalents, laws and illustrations, rules and examples, tools and their use, and the like. The matching test is a very convenient form of exercise for measuring such learning. In the words of Lindquist, "The matching exercise is particularly well adapted to testing in *who, what, when,* and *where* types of situations, or for naming and identifying abilities." [60]

Its principal limitations are as follows: (1) It is not well adapted to the measurement of understanding as distinguished from mere memory; (2) With the exception of the true-false test, the matching test is the form most likely to include irrelevant clues to the correct response; and (3) Unless skillfully made, it is time-consuming for the pupil. The suggestions that follow are designed to overcome the last two limitations. The matching test can hardly be designed to measure genuine understanding of a high level or the ability to interpret complex relationships.

I. Illustrations of Matching Tests

The following examples from well-known standard tests illustrate different mechanical arrangements of matching tests in a variety of subjects.

Every Pupil Test in Physics, 1930 Series [61]

DIRECTIONS: Read each definition or description. Then select from the Answer List the word thus defined and write its number on the dotted line in front of the definition. The answer to the sample is (Power), so 18 is written on the dotted line.

---

[60] Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *op. cit.*, page 150.

[61] Devised by F. W. Brown and others, and published by the Ohio State Department of Education.

ANSWER LIST: (Arranged alphabetically)

1. Adhesion
2. Centrifugal
3. Centripetal
4. Cohesion
5. Conduction
6. Conductor
7. Convection
8. Density
9. Efficiency
10. Energy
11. Heat of Fusion
12. Heat of Vaporization
13. Inertia
14. Insulator
15. Kinetic
16. Mechanical Advantage
17. Potential
18. Power
19. Radiation
20. Relative Humidity
21. Specific Gravity
22. Specific Heat
23. Surface Tension
24. Work

..18. SAMPLE: The rate of doing work.

..... 1. Weight per unit volume.
..... 2. Mutual force of attraction between like molecules.
..... 3. Tendency of a body to resist any change in its state of rest or motion.
..... 4. Tendency of surface of a liquid to contract as much as possible.
..... 5. Capacity for doing work.
..... 6. The ratio of resistance overcome to effort exerted.
..... 7. The product of a force and the distance through which it acts.
..... 8. Ratio of output to input.
..... 9. The energy a body possesses because of its position.
.....10. The number of calories required to melt one gram of a substance.
.....11. Amount of water-vapor the air holds compared to what it could hold at the same temperature.
.....12. Transfer of heat from a hot to a cold body by molecular collision.
.....13. Transfer of heat by means of ether waves.
.....14. The force pulling the body toward the centre of rotation.
.....15. A substance that conducts heat or electricity very poorly or almost not at all.

Cooperative Test of Social Studies Abilities, Experimental Form Q[62]

Directions: In which of the sources listed in the left-hand column would you look first to find the items listed in the right-hand column? Consider each group separately. Put the **number** of the **best** source in the parentheses after each item.

1 **Atlas**
2 *Current History*
3 **Dictionary**
4 **Economics textbook**
5 **Encyclopedia**

51. A discussion of an important present-day issue in Congress .... 51( )
52. The location of the ten largest cities in the world .... 52( )
53. How to hyphenate the word *cinema* .. 53( )
54. Amendments to the Constitution .... 54( )
55. A discussion of standards of living ... 55( )
56. The population of a particular small town .... 56( )

----

1 **American history textbook**
2 **Book of quotations**
3 **Library catalog**

57. List of news dispatches on CCC activities .... 57( )
58. A short account of the early history of Manhattan Island .... 58( )

---

[62] Devised by J. Wayne Wrightstone, and published by Cooperative Test Service.

4 *National Geographic Magazine*	**59.** The author of *Weather in the Street*	59( )
5 *New York Times Index*	**60.** Information about the growth of slavery in the United States	60( )
	61. Who said, "Brevity is the soul of wit."	61( )
	62. Pictures and story of recent developments in the TVA	62( )

1 **Daily newspaper**	**63.** The Pulitzer Prize awards of 1930 ..	63( )
2 *Readers' Guide to Periodical Literature*	**64.** Today's price quotations on stocks and bonds	64( )
3 *Time*		
4 *World Almanac*		
5 **Library catalog**		

Cooperative French Test, Junior Form, 1936[63]

Directions: Each of the English sentences and phrases below is followed by a translation in which there is a blank indicated in this way: (___). The translation will be correct when one of the five numbered words, phrases, or endings listed at the left of the group is inserted in the blank (___). Decide which of the five items will make the translation complete and correct, and put its number in the parentheses at the right-hand edge of the page.

IV

1. ce	29. These books.	(___) livres	( )
2. ces			
3. cet	30. That school.	(___) école	( )
4. celles			
5. cette	31. That money.	(___) argent	( )

VIII

1. qui	38. What are they asking for?	(___) demandent-ils?	( )
2. quoi			
3. quelles	39. Who came down the first?	(___) est descendu le premier?	( )
4. que			
5. qu'	40. Which roads are the best?	(___) routes sont les meilleures?	( )

XIII

1. -se	50. They lighted several fires.	On a allumé plusieurs feu-(___)	( )
2. -es			
3. -x	51. I didn't buy the other books.	Je n'ai pas acheté les autr-(___) livres	( )
4. -s			
5. No ending needed	52. He had black hair.	Il avait les cheveu- (___) noirs	( )

---

[63] Devised by Jacob Greenberg and Geraldine Spaulding, and published by Cooperative Test Service.

Sones-Harry High School Achievement Test [64]

SECTION G. [MATHEMATICS]

IMPORTANT THEOREMS IN GEOMETRY

Directions: In the parentheses after each geometric condition given below in Column 2, write the number of the results in Column 1 that could be proved by it.

Column 1 (Results)

1. angles equal
2. triangles congruent
3. triangles similar
4. lines perpendicular
5. lines parallel
6. quadrilateral is a parallelogram
7. parallelogram is a rectangle
8. two arcs equal (in same or equal circles)
9. two chords equal (in same or equal circles)
10. areas of polygons equivalent

Column 2 (Conditions)

66. If two opposite sides are equal and parallel . . . . ( )66
67. If perpendicular to the same line . . ( )67
68. If the sides are proportional . . . . . ( )68
69. If they have equal arcs . . . . ( )69
70. If side-angle-side equal side-angle-side respectively . . . . ( )70
71. If they are parallelograms with equal bases and altitudes . . . . ( )71
72. If their central angles are equal . . ( )72
73. If a tangent is drawn to the radius at point of contact . . . . ( )73
74. If corresponding parts of congruent triangles . . . . ( )74
75. If one angle is a right angle . . . . . . ( )75

II. Rules and Suggestions for Construction

The purpose of the first three suggestions is to avoid irrelevant clues and that of the remaining five is to reduce the amount of time required to take the test.

1. *Include only homogeneous material in each matching exercise.* Do not mix, in a single test, such dissimilar associations as persons and events, dates and events, terms and definitions. Put short titles at the top of both columns to describe the contents accurately. For example: Column 1, *Events;* Column 2, *Dates.*

2. *Check each exercise carefully for unwarranted clues that may indicate matching pairs.* For each item ask this question: "What is the least amount of information that must be known in order to select the right response?"

3. *Avoid making the test too easy.* The difficulty of a matching exercise may be increased by including more responses than needed, and by using some of the responses more than once in the same test.

4. *One list should consist of single words, numbers, or brief phrases.* In general, the column of short terms should be on the right and contain the items from which the choice is made.

5. *The items in the response column should be arranged in systematic order.* If the list consists of dates, they should be in chronological order. For other items, alphabetical order will assist the pupils in locating the desired response. The responses in the column should then be numbered consecutively.

---

[64] Devised by W. W. D. Sones and David P. Harry, Jr.. and published by World Book Company, 1929.

6. *Indicate clearly the basis upon which matching is to be done.* This should be specified both in the directions and in the column headings. The pupil will be told to put the NUMBER of the response selected in the answer space beside the test item.

7. *The matching test should contain at least five and not more than fifteen items.* Larger lists waste time and shorter lists increase the possibility of guessing the correct response.

8. *All the items for the matching tests should be on a single page.* Turning the page back and forth in search of desired responses is both confusing and time-consuming.

SELECTED REFERENCES FOR FURTHER READING

Greene, Edward B., *Measurements of Human Behavior.* New York: The Odyssey Press, 1941, Chapter 6.

Greene, Harry A., Jorgensen, Albert N., Gerberich, J. Raymond, *Measurement and Evaluation in the Secondary School.* New York: Longmans, Green and Company, 1943, Chapter VIII.

Hawkes, Herbert E., Lindquist, E. F., and Mann, C. R., *The Construction and Use of Achievement Examinations.* Boston: Houghton Mifflin Company, 1936. Chapters II–VIII.

Lang, Albert R., *Modern Methods in Written Examinations.* Boston: Houghton Mifflin Company, 1930. Chapters VI–IX.

Lee, J. Murray, *A Guide to Measurement in Secondary Schools.* New York: D. Appleton-Century Company, 1936. Chapter XI.

Odell, C. W., *Traditional Examinations and New-Type Tests.* New York: D. Appleton-Century Company, 1928. Chapters X–XVI.

Remmers, H. H., and Gage, N. L., *Educational Measurement and Evaluation.* New York: Harper & Brothers, 1943. Chapter IX.

Rinsland, Henry Daniel, *Constructing Tests and Grading in Elementary and High School Subjects.* New York: Prentice-Hall, Inc., 1938. Chapters II–VII.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman & Company, 1929. Chapters VIII, X, XI, XII, and XIII.

Tyler, Ralph W., *Constructing Achievement Tests.* Columbus: Bureau of Educational Research, Ohio State University, 1934. 110 pages.

Weidemann, Charles C., *How to Construct the True-False Examination.* New York: Bureau of Publications, Teachers College, Columbia University, 1926. 118 pages.

CHAPTER VI

The Construction and Use of Essay Examinations

To limit the use of informal teacher-made tests to those classified as objective in type is an unwarranted restriction. The so-called traditional test or essay examination still has a legitimate place in the modern school. This chapter will consider some of the advantages and limitations of this type of test, and offer suggestions for its improvement and use.

A. Limitations of the Essay Examination

As ordinarily employed, the essay examination has certain serious limitations. It suffers in comparison with most forms of objective tests on the three important criteria of a satisfactory measuring instrument, validity, reliability, and usability.

Low validity. In the first place, the essay examination as commonly used has low validity. Several factors contribute to this condition. The limited sampling of the essay examination is often pointed out. Ruch,[1] for example, produced evidence to show that the essay called forth less than half the knowledge the average pupil actually possessed on the subject as determined by objective tests, and required twice the time to do it. The essay also includes many irrelevant factors, such as the quality of the spelling, handwriting, and English used, as well as bluffing, for which no correction formula exists. It has been suggested that the essay overrates the importance of knowing how to say a thing and underrates the importance of having something to say. In view of these limitations, the ordinary essay examination has little validity as an instrument of diagnosis.

Low reliability. In the second place, the essay examination as commonly used is low in reliability. Since short tests are less reliable than long tests, the narrow sampling afforded by essay examinations would tend to restrict its reliability. Still more serious is the subjectivity of scoring. Numerous studies have shown that teachers cannot agree with each other as to the values to be allowed examination papers of the essay type. Studies have also shown that the same teachers cannot agree with themselves on a second series of values assigned independently to the same papers. Part of this is due to different standards of marking and different weighting of the questions. Certain irrelevant factors, such as the physical and mental condition of the person marking the papers, also tend to condition the mark assigned a paper by a given teacher at any particular time. An English poet states the situation as follows: [2]

---

[1] G. M. Ruch, *The Objective or New-Type Examination,* page 54. Chicago: Scott, Foresman & Company, 1929.

'Twixt Right and Wrong the Difference is dim:
'Tis settled by the Moderator's Whim:
  Perchance the Delta on your Paper marked
Means that his Lunch has disagreed with him.

In a study [3] made at the University of West Virginia, Ashburn came to the conclusion that "the passing or failing of about 40 per cent depends, not on what they know or do not know, but on *who* reads the papers" and that "the passing or failing of about 10 per cent depends . . . on *when* the papers are read." It has been observed that the scores tend to rise as time passes, and that the values assigned tend to be greatly influenced by those allowed the paper immediately preceding. For example, one writer asserts that, "A *C* paper may be graded *B* if it is read after an illiterate theme, but if it follows an *A*+ paper, if such can be found, it seems to be of *D* caliber." [4]

That this situation is not peculiar to American education is indicated by the Examination Inquiry conducted by the International Institute of Teachers College, Columbia University.[5] In fact, one prominent authority [6] asserts that evidence showed the unreliability of essay examinations in Europe was "even more serious" than had been revealed many times in America. In support of this rather surprising conclusion, he says: "In the English studies, examiners were found to reverse their judgments almost completely when asked to mark the same papers they had scored a year before."

In fairness to the essay examination, however, it should be pointed out that many of the studies reported have been with the unimproved form of the examination given under unfavorable conditions.

---

[2] Quoted by I. L. Kandel, *Examinations and Their Substitutes in the United States,* page 28. New York: Carnegie Foundation for the Advancement of Teaching, 1936.

[3] Robert R. Ashburn, "An Experiment in the Essay-Type Question," *Journal of Experimental Education,* 7: 1–3, September, 1938.

[4] John M. Stalnaker, "The Problem of the English Examination," *Educational Record,* 17: 41, Supplement No. 10, October, 1936.

[5] Published by the Bureau of Publications, 1936.

[6] W. Carson Ryan, Jr., "The Seventh World Conference of the New Education Fellowship, II," *School and Society,* 44: 364. September 19, 1936.

Often the essay examination at its worst has been compared with an improved new-type. Under such conditions, the former is bound to show up in an unfavorable light. If new-type tests had been scored under similar conditions, without scoring rules or keys, the agreement of the scores would be less impressive. As a matter of fact, even with scoring rules and keys, the agreement among the scores on new-type tests allowed by amateur scorers is far from perfect. Under favorable conditions the agreement among scorers of essay examinations approximates that reported for objective tests. One study[7] reports that the average correlation coefficient between first and second scorings of an essay test in history by three experienced scorers was .98. Another study[8] reports that the median coefficient obtained between two independent readings of certain College Entrance Board examinations was .97. All twenty of the coefficients were above .90, with the exception of English, which was .84. It must be kept in mind, however, that these examinations were so worded as to make the scoring more objective than is usually possible with ordinary essay examinations.

It should be noted that most studies having to do with the reliability of essay examinations really show the reliability of *marking the examination* rather than the reliability of the *examination* itself. A few studies have been reported of the correlation between two forms of an essay examination designed for a particular purpose which were given to the same pupils and carefully marked by experienced examiners. McGregor and Ruch[9] used this procedure in studying eighth-grade examinations in sixteen subjects from 952 pupils in eleven states. Each paper in the two sets of examinations was marked independently by two experienced teachers. This study made it possible to compare the reliability of the *examination* with the reliability of *marking the examination.* The agreement of the two independent markings of the same papers is represented by an average correlation of .62, while the agreement of the two sets of examinations marked by the same teacher is represented by an average correlation of only .43. Gordon[10] made a similar study of the New York Regents' Examinations with startlingly comparable results. Gordon found the average agreement of the two independent markings of the same papers was .72, while the average agreement of the two sets of examinations marked by the same teacher was only .42. Another study[11] conducted at the University of Chicago High School showed that two independent sets of marks assigned by two "experienced readers of essay examinations" agreed to the extent of .944 on Form A and .845 on Form B, but that the correlation between Form A and Form B was only .60. These three studies seem to agree on one important point: *The reliability of marking the essay examination is higher than the reliability of the examination itself.*

---

[7] Roy E. Cochran and Charles C. Weidemann, "Improvement of Consistency of Scoring the 'Explain' and 'Discuss' Essay Examination," a paper read before Section C of the American Educational Research Association at Cleveland, Ohio, March 1, 1939.

[8] John M. Stalnaker, "Essay Examinations Reliably Read," *School and Society,* 46: 671–672, November 20, 1937.

[9] G. M. Ruch, *The Objective or New-Type Examination,* pages 91–96. Chicago: Scott Foresman & Company, 1929.

[10] *Ibid.,* pages 97–98.

Low usability. The essay examination also ranks low in usability. There seems no escape from the fact that this type of examination is time-consuming, both for the pupil and for the teacher. In fact, the additional expenditure of time and energy over that needed for objective tests is so serious a limitation that the use of essay examinations can be justified only if it can be shown that the values realized are commensurate with this investment.

B. Advantages of the Essay Examination

Reliability and usability. Even the most enthusiastic advocate of essay examinations would scarcely claim their superiority over objective tests on the grounds of reliability or usability. The best that can be hoped for essay examinations is that by the use of improved techniques their reliability may approach that of objective tests. As regards usability, the fact that the questions can be written on the blackboard is an advantage only in those schools which lack duplicating facilities. The reduction in time required to prepare essay examinations is more apparent than real, if the work is well done. Whatever advantage arises therefrom is more than offset by such considerations as the extra time demanded for giving and scoring.

Validity. It is apparent that if the use of essay examinations can be justified it must be upon the ground of their superior validity for certain purposes. What, then, are the unique functions of these examinations?

Unfortunately, upon this crucial issue little experimental evidence exists. One study[12] indicated that about 30 to 40 per cent of the mental functions measured by improved essay tests of the "compare and contrast" type were not measured by true-false tests covering the same material. Two similar studies by Cochran and Weidemann[13] compared one-word fact tests and essay tests of the improved "explain" and "discuss" types covering the same material, and concluded that about 40 per cent of the mental functions measured by the latter were not measured by the former. The important question of just what unique mental functions each type of test measures remains to be answered.

---

[11] Arthur E. Traxler and Harold A. Anderson, "The Reliability of an Essay Examination in English," *School Review,* 43: 534–539, September, 1935.

[12] C. C. Weidemann and Lyndall Fisher Newens, "Does the 'Compare-and-Contrast' Essay Test Measure the Same Mental Functions as the True-False Test?" *Journal of General Psychology,* 9: 430–449, October, 1933.

In the absence of experimental evidence, it is necessary to fall back on logical considerations. Raths argues that the essay test is useful in measuring four of the eight objectives of instruction: namely, functional information, certain aspects of thinking, study skills and work habits, and a functioning social philosophy. It will be noted that these objectives emphasize the *functioning,* rather than the mere possession, of knowledge.

There would appear to be little justification for using essay tests for the recall of knowledge in piecemeal fashion. Sims,[14] however, analyzed 458 questions ordinarily classified as of the essay type, and found that fewer than half in the high school and fewer than one in five in the elementary school involved discussion, the others being almost equally divided between simple-recall and short-answer questions requiring not more than one sentence for a response. The Evaluation Committee of the Seattle Schools[15] came to the conclusion that the evaluation of growth in language ability would require the use of several types of tests.

> Both objective and essay tests appeared necessary to measure achievement at the various levels of knowledge. Objective tests of the multiple-choice and matching type could be used to measure achievement at the *recognition* and *recall* levels. However, evaluating achievement at the level of *interpretation* and *evaluation* would require essay-type tests, as well as certain kinds of objective tests; evaluating achievement at the level of *application* would seem to be done most effectively by essay tests, since this would involve measuring the student's ability to utilize information learned in one situation in the solution of problems in a new setting.

One other advantage of the essay examination should be mentioned. Several experimental studies[16] have shown that the type of measurement used by the teacher influences the type of study procedures employed by the pupils. When pupils expect the test to be of the essay type, in whole or in part, they seem more likely to employ such desirable study techniques as making outlines and summaries, and seeking to perceive relationships and trends, than is done when objective tests are used exclusively.

---

[13] See *Phi Delta Kappan,* 17: 59–61, 75, December, 1934; and 19: 113–115, 131, January, 1937.

[14] Verner Martin Sims, "Essay Examination Questions Classified on the Basis of Objectivity," *School and Society,* 35: 100–102, January 16, 1932.

[15] Helen F. Olson, "Evaluating Growth in Language Ability," *Journal of Educational Research,* 39: 247, December, 1945.

[16] For a fuller discussion of this point, see pp. 339–343.

The practical conclusion follows that neither the essay nor objective test should be used exclusively. From Lee and Segel's[17] analysis of the measurement practices of 1,600 secondary school teachers, distributed widely over the United States, it appears that two thirds of the teachers favor the use of a combination of the two types. It is encouraging that the practice of more and more teachers seems to be governed by the sound philosophy of measurement stated by Lindquist in the following sentence:[18]

> The intelligent point of view is that which recognizes that whatever advantages either type may have are *specific* advantages in *specific* situations; that while certain purposes may be best served by one type, other purposes are best served by the other; and, above all, that the adequacy of either type in any specific situation is much more dependent upon the ingenuity and intelligence with which the test is *used* than upon any *inherent* characteristic or limitation of the *type* employed.

C. Suggestions for Improving Essay Examinations

Although the essay examination has been in existence for hundreds of years, the amount of research devoted to it is much less than that devoted to the objective test, which is comparatively new. Furthermore, practically all the research relating to the former has been of a negative kind. Its purpose has been to show how poor unimproved essay examinations are, rather than to devise means for their betterment. However, a study of the meager experimental literature does yield several positive suggestions. The next two sections will be devoted to a consideration of some of the most promising of these suggestions.

Improving the construction and use of essay examinations. It is just as important to know *where* to use the essay examination as it is to know *how* to use it. It is wise to restrict the use of the essay test to the measurement of those functions for which it is best adapted. There would usually appear to be no good reason for employing subjective measurement where objective measurement is adequate. What, then, does the essay examination attempt to do?

Weidemann[19] recognizes eleven definable types of improved essay examinations. Arranged in a series from simple to complex, these types are as follows: (1) *what, who, when, which,* and *where;* (2) *list;* (3) *outline;* (4) *describe;* (5) *contrast;* (6) *compare;* (7) *explain;* (8) *discuss;* (9) *develop;* (10) *summarize;* and (11) *evaluate.* The first two types seem hardly distinguishable from recall tests of the objective type. Ten years earlier Monroe and Carter[20] made a very suggestive classification of thought questions into twenty types. These types, together with an illustration of each, taken from the field of measurement, appear below.

---

[17] J. Murray Lee and David Segel, *Testing Practices of High-School Teachers,* page 38. United States Office of Education Bulletin, No. 9, 1936.

[18] Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *The Construction and Use of Achievement Examinations,* page 20. Boston: Houghton Mifflin Company, 1936.

[19] C. C. Weidemann, "Written Examination Procedures," *Phi Delta Kappan,* 16: 78–83, October, 1933; also, C. C. Wiedemann, "Review of Essay Test Studies," *Journal of Higher Education,* 12: 41–44, January, 1941.

Thought Questions

1. Selective recall—basis given.
   Name three important developments in measurement which occurred during the first decade of the twentieth century.
2. Evaluating recall—basis given.
   Name the three persons who have had the greatest influence on the development of intelligence testing.
3. Comparison of two things—on a single designated basis.
   Compare essay examinations and objective tests from the standpoint of their effect upon the study procedures used by the learner.
4. Comparison of two things—in general.
   Compare standardized and non-standardized tests.
5. Decision—for or against.
   In which, in your opinion, can you do better, oral or written examinations? Why?
6. Cause or effects.
   How do you account for the great popularity of objective tests during the last twenty-five years?
7. Explanation of the use or exact meaning of some phrase or statement in a passage.
   What is the meaning of "Delta" in the verse quoted on page 166?
8. Summary of some unit of the text or of some article read.
   Summarize in not more than one page the advantages and limitations of essay examinations.
9. Analysis (The word itself is seldom involved in the question).
   Why are so-called "progressive educators" suspicious of standardized tests?
10. Statement of relationships.
    Why is it that all essay examinations, regardless of the school subject, tend to be measures of the learner's mastery of English?
11. Illustrations or examples (your own) of principles in science, construction in language, etc.
    Give two original examples of specific determiners in objective tests.

---

[20] Walter S. Monroe and Ralph E. Carter, *The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students,* 26 pages. Urbana, Illinois: Bureau of Educational Research Bulletin, Number 14, University of Illinois, 1923.

12. Classification (usually the converse of No. 11).
    What type of error appears in this test item? "With what Balkan country did the Allies fight in World War I?"

13. Application of rules or principles to new situations.
    In the light of China's experience with state examinations what would you expect to be the effect of the Regents' Examinations in New York?

14. Discussion.
    Discuss the place of measurement in science.

15. Statement of aim—author's purpose in his selection or organization of material.
    In view of the author's discussion on page 21, why are so many authorities quoted in Chapter I?

16. Criticism—as to the adequacy, correctness, or relevancy of a printed statement, or a classmate's answer to a question on the lesson.
    Criticize or defend the statement, "The essay examination overrates the importance of knowing how to say a thing and underrates the importance of having something to say."

17. Outline.
    Outline the principal steps in the construction of an informal teacher-made test.

18. Reorganization of facts (a good type of review question to give training in organization).
    Name ten practical suggestions from Chapters IV, V, and VI that are particularly applicable to the subject you teach or plan to teach.

19. Formulation of new questions—problems and questions raised.
    What are some problems relating to the use of essay examinations that require further study?

20. New methods of procedure.
    Suggest a plan for proving the truth or falsity of the contention that exemption from examinations is a good policy in high school.

It will be noted that the classifications by Weidemann and by Monroe and Carter recognize a considerable number of rather distinct abilities, which are measurable by essay tests. It is probably best to measure each one separately rather than to attempt to measure several of them by the same test. It will be further noted that the emphasis in most of these types is upon organization, relationship, evaluation, application, or some similar ability to which a purely objective test is poorly adapted. Teachers should study carefully each type of essay question until they are familiar with its distinguishing characteristics. If a proposed essay question does not seem to conform to one of these types, it had usually better be reworded or adapted to some form of the objective test. No question should be included until its purpose has been clearly defined.

The essay examination would appear to be particularly valid in two situations. The first of these is obviously in such courses as

English composition and journalism, where the student's ability to express himself effectively is the major objective of instruction. The second situation is in advanced courses of other subjects, where critical evaluation and the ability to assimilate and organize large amounts of material constitute important objectives. In this connection it is significant to note that Jones[21] found that 68 per cent of the college students who took senior comprehensive examinations and 55 per cent of the superior students in other colleges stated their views as follows: "I think one's ability is far better shown through discussion questions than through short objective questions."

There is some evidence that a more valid sampling of the pupil's knowledge is afforded by increasing the number of questions and reducing the length of discussion expected on each. In many cases a well-constructed paragraph is sufficient. Very few discussions need exceed one or two pages in length. In any case, the question should be so worded as to restrict the responses toward the objective which it is desired to measure. For example, Wrightstone suggests that the question, "Explain the reasons for the strike at General Motors in 1937," is too general, and would be improved if it were restricted by the addition of the phrases "to show (a) the labor grievances of the employees; (b) the practices of the employer; (c) related national, social, and economic factors; (d) the rival labor unions; and (e) the method of striking."[22] It must be recognized, however, that such suggestions, at least in part, take from the essay examination its uniqueness. The proposed modifications may appear to improve the reliability of the traditional examination by the obvious device of making it more like the objective test.

One of the difficulties with constructing essay tests is that the process appears so easy. As a matter of fact, it is probably more difficult to construct essay tests of high quality than it is to construct objective tests of high quality. Much care and thought must be given to their construction, if tests of any kind are to measure anything but mere memory for factual knowledge. Many of the general principles of testing outlined in an earlier chapter are as applicable to essay tests as to objective tests. There is always risk that, in attempting to phrase essay questions so that they can be scored more objectively, the result may be made less satisfactory than an out-and-out objective test. Critical revision, utilizing, if possible, the judgment of a colleague, is especially important.

Some writers have emphasized the importance of training pupils in taking examinations. Worcester[23] suggests that the essay examination is "obviously invalid and unfair," at least in part, because the pupils are being required to take a test on a type of work for which they have had no specific training. The rational solution offered is to supply the necessary training rather than to abandon the essay examination. Wider experience and training in preparing for and in taking tests of all types is likely to increase the accuracy of measurement. Edmiston[24] prepared instructions to pupils for taking examinations which were far more elaborate than the usual directions accompanying tests. He found that the use of these instructions increased the validity of the examinations and produced "definite improvements in students' records of achievement from examinations." It would appear wise to provide instruction of this sort in the regular program of studies. It is most unfortunate when a pupil fails to receive recognition for knowledge he actually possesses simply because he has not mastered the technique of putting it on paper. Edmiston's suggestions,[25] given below, will prove helpful in planning such a program of instruction.

---

[21] Edward Safford Jones, *Comprehensive Examinations in American Colleges,* page 373. New York: The Macmillan Company, 1933.

[22] J. W. Wrightstone, "Are Essay Examinations Obsolete?" *Social Education,* **1**: 403, September, 1937.

Important Considerations in Taking Examinations

1. Your name should appear on the first or last sheet of the examination, if sheets are securely bound. Each loose sheet should have the name entered inconspicuously, preferably on back *where it will not be seen by the scorer, when scoring.*
2. Write legibly. Your answer can't be right if it can't be read. If a *T* cannot be distinguished from an *F*, the answer is wrong. Be sure your pen or pencil (if allowed) fosters distinct and not blurred writing.
3. Use terms or a vocabulary suited to the subject. Do not use a word unless its meaning is clear to you, and repeat a word rather than use another which may not have exactly the same desired meaning.
4. Space (the back of sheets, the margins, or an extra sheet) should be used for
   - *a.* computations.
   - *b.* practice in the formation of desirable statements, not padded but furnishing quality rather than quantity to the answer.
   - *c.* the hasty jotting of facts pertaining to some questions when these facts arise, while working upon another question.
5. *The statement of each question must be fully considered.* Carelessness not only penalizes the student but also lowers the dependability of the measurement obtained by the instructor.

---

[23] D. A. Worcester, "On the Validity of Testing," *School Review,* 42: 527–531, September, 1934.

[24] R. W. Edmiston, "Examine the Examination." *Journal of Educational Psychology,* 30: 126–138, February, 1939.

[25] *Ibid.,* pages 137–138.

6. The directions telling how to answer the questions should be carefully followed. Underscore the important points in the directions.
7. In essay questions, *underscore* the part of the statement that furnishes the direct question asked. Then underscore any parts of the statement which furnish data for the answer. Number each part so that you will not omit anything from your answer.
8. Proceed directly through the examination with no lengthy consideration of unfamiliar points. After completing the parts which were readily answered, start again and answer those questions which yield to more diligent effort. Do not waste time by trial and error method upon questions which bring no recognition or recall of related materials. After completing the second consideration of the test, spend the remainder of the time upon the more familiar of the unanswered questions. Note that hesitation wastes time, ruins confidence, and destroys mind-set.
9. If after thorough consideration you do not understand some direction or question due to other than lack of knowledge of the course, call the attention of the person in charge with as little disturbance as possible in order that the tester may come to your seat or allow you to come to him as conditions may determine.
10. Reread each answer before passing to the next question and the completed examination before delivery to the instructor. Is the meaning clear and writing legible?

By way of summary, three important suggestions for the construction and use of essay examinations are as follows:

1. Restrict the use of the essay examination to those functions to which it is best adapted. When it is not clear that the essay type is required for measuring the desired objective, use the objective test.
2. Increase the number of questions asked and reduce the amount of discussion required on each. Always indicate clearly the type of discussion desired.
3. Make definite provisions for teaching pupils how to take examinations. Specific training in preparing for and in taking tests and examinations of the various types commonly encountered is a legitimate objective of instruction.

Improving the scoring or grading of essay examinations. Rinsland [26] makes a distinction between the terms *scoring* and *grading*. Scoring is an objective process of counting right or wrong responses, whereas grading always means interpreting quality in terms of some criterion. Strictly speaking, then, it is more correct to speak of grading or rating essay examinations than it is to speak of scoring them.

It is, of course, apparent that whatever claims are made for the validity of the essay test as a measuring instrument are conditioned upon the assumption that the papers can be read accurately. Not only must the essay test, for example, *call forth* from superior pupils responses which are consistently superior, but the teachers marking the papers must be able consistently to *recognize* that they are superior responses. The same is true of responses with other degrees of merit. The grading of the essay examination, therefore, occupies a strategic position.

---

[26] Henry Daniel Rinsland, *Constructing Tests and Grading in Elementary and High School Subjects,* page 302. New York: Prentice-Hall, Inc., 1938.

To begin with, certain preventive measures are important. A careful wording of the questions, and directions to the pupil which indicate clearly just what type of response is expected will simplify the problem of marking the papers. The use of optional questions should be discouraged.[27] The simple precaution of having the pupil record his name inconspicuously either on the back or at the end of the paper, rather than at the top of each page, is likely to increase the accuracy with which the paper is graded.

Cochran and Weidemann[28] outline a procedure for evaluating essay examinations, the essentials of which can be taught in ten minutes. This is shown by the fact that the majority of the consistency coefficients of two series of scorings made five weeks apart were between .80 and .90 for teachers with ten minutes of training. Independent scores by experienced readers showed an average agreement of .98 when the procedure given below in a slightly modified and abridged form was used.

SUGGESTIONS FOR MARKING ESSAY EXAMINATIONS
(After Cochran and Weidemann)

1. I read over a sampling of the papers to obtain a general idea of the grade of answer I may expect.
2. I score one question through all of the papers before I consider another question. I have found two outstanding advantages in scoring one question through an entire set of papers. The first is that the comparison of answers appears to make the grades more exact and just. The second is that having to keep only one list of points in mind saves time and promotes accuracy.
3. Before scoring any papers I read the material in the text which covers the questions, and also the lecture notes on the subject.
4. I make a list of the main points which should be discussed in every answer. Each of these points must be weighed and assigned a certain value if the scoring is to approach accuracy. This value assigned to the main points needed for a reasonably adequate answer is designated as the minimum score. If a pupil elaborates and discusses points not required yet pertinent to the question, his answer is given an additional value, called the extra score. This extra score may vary for different pupils, but may not exceed a certain set maximum.
5. After the points have been weighed, the actual scoring begins. I read the

---

[27] John M. Stalnaker, "A Study of Optional Questions on Examinations," *School and Society,* 44: 829–832, December 19, 1936.

[28] Roy E. Cochran and C. C. Weidemann, *op. cit.*

> answer through once and then check back over it for fact details. I attempt to mark every historical mistake on the paper and write in briefly the correction. As I read the answer I make a mental note of the points omitted and the value of each point, so that when the end of the question is reached, I have the minimum grade figured. If there is any additional or extra percentage to be given, it is added to the minimum score, and then the value of the question is written in terms of the per cent deducted rather than the positive per cent. Then when every question on a paper is scored, it is a simple matter to add the negative quantities and obtain the final grade.

It is difficult to overemphasize the importance of three things: (1) the preparing in advance of a list of answers which are considered adequate for the objectives of the test; (2) the assigning of a specific value to each essential part of the answers; and (3) the grading of one question through all the papers before going on to another question. Most students of the problem recommended attempting to distinguish a relatively small number of degrees of merit in an answer. Perhaps as good a plan as any is to allow credit for each part of the answer considered essential to a question as follows: 3 for superior, 2 for average, 1 for inferior, and 0 for an omission or wrong reply. Stalnaker[29] found that the weighting of essay questions was of negligible value—the correlation between weighted and unweighted scores on the College Entrance Board Examinations varying from .97 to .997.

In addition to the points made by Cochran and Weidemann, several authorities have found another suggestion helpful. The suggestion is to make a sorting of the papers into three to five piles, according to the merit of the discussion of each question on the basis of a brief preliminary examination of the answers. Sims[30] describes clearly a very satisfactory procedure as follows:

1. Quickly read through the papers and on the basis of your opinion of their worth sort them into five groups as follows: (a) very superior papers, (b) superior papers, (c) average papers, (d) inferior papers, (e) very inferior papers. (Remember that in a normal group you would expect to have approximately 10 per cent of *very superior* and 10 per cent of *very inferior* papers, 20 per cent of *superior* and 20 per cent of *inferior* papers, and 40 per cent of *average* papers. Do not, however, try to conform rigidly to this rule. Your group may not be a normal one.)
2. Reread the papers in each group and shift any that you feel have been misplaced.

It will be noted that Sims does not suggest a rigid distribution according to the normal curve. A general understanding of the theory of probability is desirable because it underlies the measurement of human ability and achievement at many points. In classes for which satisfactory evidence exists as to their non-typical character, *skewed* distributions may be regarded as *normal for those particular classes.* On the other hand, any consistent skewness in the marks on an examination can be justified only on the ground that the class is definitely superior or definitely inferior, or else that the examination is too easy or too difficult.

---

[29] John M. Stalnaker, "Weighting Questions in the Essay-Type Examination," *Journal of Educational Psychology,* 29: 481–490, October, 1938.

[30] Verner Martin Sims, "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating," *Journal of Educational Research,* 24: 216–223, October, 1931.

The preliminary sorting of the papers into piles of approximately equal merit before assigning numerical values to them will help to avoid the difficulty pointed out by Stalnaker: namely, that the values allowed a paper are often greatly influenced by the merit of the paper which happens immediately to precede it in the order of scoring. It is also easier to locate papers distinctly out of line with those in a particular group supposedly of similar quality. It is a good idea to throw the papers into a single group after each question has been evaluated and before they are re-sorted into piles according to the merits of the discussions of the next question. This procedure will make it easier to conceal the identity of the particular pupil whose paper is being judged and so to avoid one of the most disturbing factors in marking essay examinations.

The school should adopt a policy regarding what factors shall be considered, and what factors shall not be considered, in evaluating a written examination. *Only those factors should be taken into account which afford evidence of the degree to which the pupil has attained the objectives set up for that particular course.* Except in English classes, this will rule out making arbitrary reductions for such things as faulty sentence structure, paragraphing, handwriting, and the spelling of nontechnical words. These factors will be considered only in so far as they affect the clarity of the pupil's discussion. It is always legitimate to hold the pupil responsible for the spelling, as well as the meaning, of the vocabulary which is peculiar to the course.

This does not mean that the quality of the written English used in examinations is unimportant and should therefore be disregarded. On the contrary, it is always very important. But it should be considered only in relation to that for which it may be accepted as valid evidence: namely, in determination of the pupil's mark in English. Where the teacher has complete charge of an entire grade, this adjustment is easy to make. But where the school is departmentalized the problem is more difficult. Even here it should be possible to work out a system whereby at intervals the papers in other subjects, after having been graded as to content, may be turned over to the English teacher to be judged from the viewpoint of their merits as

English compositions. In this way it may be possible to sample the pupil's characteristic performance in written English better than when he writes a paper specifically for the English teacher. And, what is equally important, it makes the pupil's mark in other subjects solely a measure of achievement in those subjects, rather than partly a measure of skill in English composition.

SELECTED REFERENCES FOR FURTHER READING

Greene, Harry A., Jorgensen, Albert N., and Gerberich, J. Raymond, *Measurement and Evaluation in the Secondary School.* New York: Longmans, Green & Company, 1943. Chapter VII.

Lang, Albert R., *Modern Methods in Written Examinations.* Boston: Houghton Mifflin Company, 1930. Chapter IV.

Monroe, Walter S., and Carter, Ralph E., *The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students.* Bureau of Educational Research Bulletin, Number 14. Urbana: University of Illinois, 1923. 26 pages.

Odell, C. W., *Traditional Examinations and New-Type Tests.* New York: D. Appleton-Century Company, 1928. Chapters VII and VIII.

Remmers, H. H., and Gage, N. L., *Educational Measurement and Evaluation.* New York: Harper & Brothers, 1943. Chapters VIII and XII.

PART III

THE TESTING PROGRAM

CHAPTER VII

Steps in the Testing Program

Dear Professor:

I have decided to give some tests in my school this fall. Please suggest a few good ones I might try. Also let me know where to get them and what they will cost.

Dear Professor:

We gave the Up-to-Date General Achievement Tests at the beginning of the school year. As we now have most of them scored, please advise me how to use the results so as to get the most good out of them. Any help will be greatly appreciated.

Probably every college professor who offers courses in measurement has received letters like those above. They indicate that some school is undertaking, or has already undertaken, to use standard tests without understanding what it is all about. Always, testing should have a *program* to guide it. What, then, is a "testing program"?

General considerations. The word "program" has certain important implications, such as *order, system, planning.* It implies a sequence of events that has been determined upon after careful thought, rather than some haphazard, hit-or-miss affair. One of the chief weaknesses of many attempts to use standard tests is that there has been no program worthy of the name. The whole procedure has simply led a precarious hand-to-mouth existence from beginning to end.

Spence [1] has suggested that "a good testing program should be supplementary *not* duplicative, usable *not* confusing, economical *not* burdensome, comprehensive, *not* sporadic, suggestive *not* dogmatic, progressive *not* static." Such a program, at least in tentative form, may very well cover an extended period, rather than be adopted piecemeal year by year. One advantage of this long-range planning is that it makes possible a varied program without leaving gaps or involving needless duplication. Stenquist,[2] speaking from wide experience, strongly advocates "some sort of systematically recurring schedules as opposed to sporadic testing," since schedules make possible "enormously greater gains" from testing. Spence offers for elementary schools what he calls "a conservative approach to the problem." This program is given in Table 7.

---

[1] Ralph B. Spence, "A Comprehensive Testing Program for Elementary Schools," *Teachers College Record,* 34: 279–284, January, 1933.

[2] John L. Stenquist, "Recent Developments in the Uses of Tests," *Review of Educational Research,* 3: 60, February, 1933.

TABLE 7

PLAN FOR A TESTING PROGRAM FOR THE ELEMENTARY SCHOOL (AFTER SPENCE)

GRADE	INTELLIGENCE [a]	ACHIEVEMENT BATTERY ANNUAL (GIVEN IN MAR. OR APR.) [b]	ACHIEVEMENT TESTS FOR SPECIAL EMPHASIS ALL GRADES FROM 3 OR 4 TO 8 ROTATING (GIVEN IN OCTOBER) [b]
Kdg.—I	Two group tests		
II		Reading Battery	
III		Skill Subjects Battery	First Year—Reading
IV	One Group test	Complete Battery	Second Year—Arithmetic
V			Third Year—Social Studies
VI		Complete Battery	
VII			Fourth Year—Language Usage and Spelling
VIII			
		Complete Battery	Fifth Year—Reading, etc.

[a] Retests for special cases as needed.

[b] All dates based on groups beginning a grade in September. Change to corresponding dates for groups beginning a grade in February. Teachers use diagnostic tests throughout the year.

It will be noted that this program calls for the use of both intelligence tests and achievement tests, and for the use of test batteries as well as of tests in the separate subjects. It is also expected that the program will merely supplement rather than supplant the ordinary informal tests and examinations made by the classroom teacher. A slight modification of the schedule as presented would involve giving a general test battery in all subjects about every third year, and an intensive program limited to one subject in each of the intervening years. The cost of such a program of standard tests would be from ten to fifteen cents per pupil per year. If the tests are intelligently used, it is doubtful whether greater returns can be had by the school from the same amount of money spent in any other way.

Traxler,[3] in giving a practical discussion of the planning and administration of the testing program, divides tests into two broad categories. The first includes group tests of general intelligence and achievement tests in the major subject-matter areas. These should be administered at regular intervals to every normal pupil in the school. The second category includes individual intelligence tests, special aptitude tests, personality tests, and tests of vocational interests.

---

[3] Arthur E. Traxler, "Planning and Administering a Testing Program," *School Review,* 48: 253–267, April, 1940.

The following comprehensive "Platform for the Use of Standard Tests" has been prepared by a committee of Massachusetts teachers: [4]

1. Scientific measuring instruments and the scientific method are badly needed in present educational practice. No business of the financial magnitude of education spends so little time and money for objective and scientific fact finding.
2. Standardized tests and measurements can fulfill their function of giving direction and efficiency to education only when used intelligently by teachers and administrators who have kept abreast of current knowledge on the subject and who are willing to follow the authors' directions for the administration and scoring of the tests used. The results of tests in which directions are not followed are worse than useless; they are misleading.
3. Every standardized test administered should be given for a specific purpose, and having been given, its results should be used. Tests which are administered, scored, and piled in a cupboard serve no useful purpose.
4. Standardized tests can be used most efficiently when their use is planned over a long period of time.
5. Standardized tests have furnished valuable information to the school administrator in practically every instance in which they have been used. The possibilities of diagnostic tests in improving instruction through analysis and diagnosis of individual and class weaknesses have not nearly been realized. Tests are of the greatest value when their results cause a teacher to redefine his objectives, alter his methods, and redirect his emphasis as a result of new, increased, and more exact knowledge about his pupils.
6. If standardized test results are to be used in measuring the efficiency of instruction, the conditions of scientific experimentation must prevail with all contributing factors defined, measured, and controlled. Failure to observe these conditions often results in teaching for test results alone, which not only invalidates any results which may be obtained, but also neglects some of the most desirable outcomes of good teaching which cannot be measured by tests. On the other hand, standardized test results cannot be ignored. They can be of great help to an administrator in judging a teacher's work, but they cannot be used as a substitute for classroom visiting, supervision, and critical subjective analysis.
7. No important decision regarding the placement of an individual pupil should be made on the basis of the result of one test of any kind. Educational achievement, mental age, I.Q., chronological age, health, teacher's judgment, physical development, social age, and emotional maturity are all factors to be considered in individual placement or any plan for grouping.
8. The content or items of a standardized test should never be used as material for class presentation and drill either before or after the administration of the

---

[4] "The Use of Standardized Tests in Massachusetts," *Test Service Bulletin No. 38,* published by World Book Company, 1938.

test. To reproduce any part of the test, either on paper or the blackboard, is not only a violation of the publisher's copyright, but will invalidate that test for future use in the school. For this reason, all copies of standardized tests should be accounted for, and extra copies should not ordinarily be left in the hands of the classroom teacher.

9. The I.Q. or mental age obtained from one group test of intelligence is much less reliable than an average of I.Q.'s or mental ages obtained from the results of two or more group tests of intelligence. An individual test of intelligence is more valid and reliable than group tests only when it is administered by a skillful and well-trained psychometrist.
10. The use of standardized tests and a knowledge of the methods used in their construction should result in an improvement in teacher-made measures of achievement.

One must not assume that the testing program should be restricted to the use of standard tests. As has been explained in the three preceding chapters, informal or teacher-made tests will have a large place in any complete testing program. Schools should have a carefully thought out general policy on such matters as the frequency of testing, the importance of final examinations, the factors to be considered in determining final marks, and, most important of all, the uses to be made of the results.

Regardless of its scope, the complete testing program at any particular time will ordinarily consist of the following eight steps, or stages, in chronological order:

1. Determining the purpose of the program.
2. Selecting the appropriate test or tests.
3. Administering the tests.
4. Scoring the tests.
5. Analyzing and interpreting the scores.
6. Applying the results.
7. Retesting to determine the success of the program.
8. Making suitable records and reports.

A. Determining the Purpose of the Program

It must be recognized at all times that tests are only tools, and that measurement is always a means to an end, never an end itself. In the final analysis, then, the value of any testing program depends upon the use made of results. Unless something is going to be done about it in the end, there is no point to beginning. Merely "giving tests" without rhyme, rule, or reason is money, time, and effort wasted. The author once heard an experienced educator say that he had wondered for years what many people did with standard tests after they had been "given." At last he found out: They filed them! The real testing program has a more serious purpose than that. The first step, therefore, in planning a program is to determine its purpose. In so doing, three things should be kept in mind.

1. It should be co-operative.
2. It should be practical.
3. It should be definite.

A co-operative program. As a rule, the program should not represent the judgment of any one person alone, but that of a group. It should be a truly co-operative enterprise. The teachers and administrative officers alike should be made to feel that it is "our" program, as, indeed, it should be. This is not likely to be the case, however, if the principal, superintendent, or research department determines the program and then "hands it down" to the classroom teachers. The entire staff should have a voice in determining the purpose of the program and in formulating the plans, and all should have the opportunity of participating in it in every way possible from beginning to end. If this is not done, the teachers are not likely fully to understand the program or to appreciate what it is attempting to do. Without the hearty co-operation of the entire staff, from the superintendent to the youngest teacher, the program is almost sure to fall short of its highest possibilities. It is suggested, therefore, that in the small school or school system the purpose of the program be decided upon after discussion in a general teachers' meeting or series of meetings in which everyone has a chance to participate. In the larger school systems it is better to entrust a committee representing all interested groups with the responsibility of planning the program. Even then it should be brought before the entire staff before final action is taken. It cannot be too strongly emphasized that the success of the program largely depends upon co-operative action. An important part of the program, therefore, is the educating of the staff so that they can participate intelligently in it. Boyer emphasizes the fact that the teacher's attitude is the most important factor to be considered in any plan, for "what she thinks and what she does as a result of her thinking, determines the success or failure of the plan." [5]

A practical program. The general purpose of the testing program is to provide data which will help in the solution of some practical school problem. As a rule, this means that the problem whose solution is sought will have to do with administration, instruction, or research, or with some combination of these three. Even when tests are used primarily for administrative purposes, such as classification, they can also be used by the classroom teachers for diagnostic purposes. Unless the school has had considerable experience with testing, it will be better not to undertake a program primarily for research, although under favorable conditions research is a legitimate interest both to classroom teachers and to administrators. Even when the program is undertaken for research purposes, the problem chosen should ordinarily be one which bears directly upon some practical issue in the school, such as determining the relative efficiency of different teaching methods or of administrative organizations.

---

[5] *Thirty-Fifth Yearbook of the National Society for the Study of Education, Part I,* page 213. Bloomington, Illinois: Public School Publishing Company, 1936.

Table 8 gives a list of ten administrative problems and ten instructional problems in the solution of which tests are useful. The list is by no means exhaustive and is merely meant to suggest the wide variety of problems that might be undertaken. In formulating the purpose of the testing program for any given situation, it is usually well to list a series of specific questions to which answers are sought. Table 8 suggests that intelligence tests have a wide usefulness in the solution of administrative problems and nonstandardized achievement tests in the solution of instructional problems. Of course, there is no clear-cut distinction between administrative and instructional problems in many instances, the same problem often having two aspects.

Table 9 shows the percentage, reported by Lee and Segel,[6] of 1,614 high-school teachers using informal objective, essay, and standardized achievement tests for 14 different purposes. This list includes all the uses mentioned by as many as 10 per cent in any one department of instruction. Perhaps the most significant single fact about this table is the remarkable agreement among the three types of tests as to the uses made of the results. The one notable exception is the use of the norms on standardized tests for comparative purposes. A recent state-wide survey of the uses of standardized tests in the secondary schools of New Jersey shows a definite shift in emphasis from marking and survey uses to the diagnosis and guidance of individual pupils.[7]

A definite program. It is not enough that the program be cooperative and practical. It must also be definite. The scope of the program may vary all the way from a single subject in one grade to a complete measurement of the entire school system. A common mistake of a staff inexperienced in the use of tests is to undertake too much. The danger then is that the program will drag along until everybody is more or less "fed-up" with it. Much of the value

---

[6] J. Murray Lee and David Segel, *Testing Practices of High-School Teachers*, pages 19–22. United States Office of Education Bulletin, No. 9, 1936.

[7] *Test Service Bulletin No. 42,* World Book Company, 1940.

TABLE 8

TYPES OF PROBLEMS FOR WHICH TESTS ARE USEFUL

TYPE OF PROBLEM	TYPE OF TEST USED			
	INTELLIGENCE		ACHIEVEMENT	
	General	*Specific*	*Standard*	*Nonstandard*
Administrative Problems:				
1. To determine whether the school is properly classified.	x		x	
2. To determine entrance requirements for grade, class, or subject.	x	x	x	
3. To form sections within the grade, special classes, etc.	x	x	x	
4. To determine intellectual level of grade, class, or school.	x			
5. To determine efficiency of class or school by comparison with norms.	x		x	
6. To determine whether proper emphasis is given various subjects.			x	
7. To stimulate interest of teachers in improvement of instruction.		x	x	
8. To assist in educational and vocational guidance.	x	x	x	x
9. To measure progress of grade or school for half year or year, etc.			x	x
10. To bring about a better understanding with parents regarding progress and plans of pupils.	x	x	x	
Instructional Problems:				
1. To motivate learning.			x	x
2. To measure class progress for a given period.			x	x
3. To provide necessary practice or drill.			x	x
4. To determine whether proper emphasis is given subjects or phases of subjects.			x	
5. To diagnose weaknesses of groups and individual pupils.			x	x
6. To determine whether pupils are working up to capacity.	x	x	x	
7. To determine status of class in relation to norms.			x	
8. To determine when class is ready for promotion.			x	x
9. To provide a basis for school marks.			x	x
10. To determine sections within class for instructional purposes.	x	x	x	x

of the information sought from the tests will be lost unless the information is made available without delay. It is usually best, particularly with inexperienced teachers to run the risk of undertaking too small a program rather than one too large.

Another mistake is in stating the purpose of the program in too general terms. "To improve instruction" is too vague and inclusive.

TABLE 9

PERCENTAGE OF 1,614 HIGH-SCHOOL TEACHERS USING INFORMAL OBJECTIVE, ESSAY, AND STANDARDIZED ACHIEVEMENT TESTS FOR VARIOUS PURPOSES (AFTER LEE AND SEGEL)

USE	INFORMAL TESTS		STANDARDIZED ACHIEVEMENT TESTS
	Objective	*Essay*	
1. To aid in determining the pupil's mark	55	30	13
2. To discuss what parts of a topic need to be re-taught	53	21	10
3. To show pupils in what part of the subject they are weak	47	20	12
4. To stimulate pupils to do better work	44	20	12
5. To discover what parts of a topic or unit need to be taught	36	12	10
6. To aid in determining which pupils will fail	34	18	10
7. To compare the results attained in two or more of my classes	31	9	14
8. To enable the teacher to tell whether poor work is due to lack of ability or other factors which can be corrected	28	14	9
9. To aid in discovering which pupils are capable of doing exceptional work	17	12	9
10. To aid in studying and advising failing pupils	15	8	8
11. To satisfy parents that their children have been marked fairly	15		8
12. To form ability groups within the room	9		
13. To furnish an estimate of the pupil's probable success in college	5		4
14. To compare the results attained by my class with the norms			20

"To motivate study" or "to diagnose weaknesses and provide a basis for remedial instruction" would be better. Best of all would be a still more definite formulation, such as "to motivate study in fifth-grade arithmetic" or "to make a diagnosis of characteristic weaknesses in first-year algebra and to formulate a program of remedial teaching to strengthen them." The purpose should state specifically both the *nature* and the *scope* of the program to be undertaken. Later chapters will discuss in some detail important administrative and instructional problems which tests may help to solve. In a long-time program the purpose for each year will have a definite relationship to the whole. No matter how stated, however, there is really one fundamental purpose in all measurement: namely, the better understanding of the individual pupil. To accomplish this purpose the information must be as definite and as complete as possible.

B. Selecting the Appropriate Test or Tests

When the purpose of the testing program has been determined, and not until then, the selection of the test, or tests, is in order. In Chapter III attention was called to the fact that a test may be superior for one purpose and worthless for another. Great care must therefore be exercised in order to secure the tests most appropriate for the purpose. Three questions require consideration:

1. Who shall select the test or tests?
2. What type of tests shall be used?
3. What is the best procedure in making the selection?

Who shall select the tests? The best qualified person, or persons, available should make the selection. In larger school systems the director of research is usually that person. But, even then, in the selection of achievement tests for specific subjects, the teachers of these subjects should be consulted, as their knowledge is essential in judging the curricular validity of the tests. In smaller schools the major responsibility is usually entrusted to the principal or superintendent.[8] However, in the selection of achievement tests a committee of teachers will be helpful in judging the content of the tests, while the principal or superintendent can pass on the technical phases of their construction. It is a sound principle in all evaluation that involves a subjective element to rely, whenever possible, upon the combined judgment of a group of competent persons rather than upon the judgment of any one individual.

What type of tests shall be used? Ordinarily an adequate testing program will involve the use of more than one type of test. It will be desirable, except in a few cases such as in the beginning of the kindergarten or first grade, to use both general intelligence and achievement tests. If considerations of time and money make it advisable to limit the testing program to one standard test for de-

---

[8] State-wide surveys in Massachusetts and New Jersey indicate this clearly. See *Test Service Bulletins No. 38* and *No. 42,* World Book Company.

TABLE 10—ADVANTAGES AND LIMITATIONS OF STANDARDIZED

CRITERION	STANDARDIZED	
	Advantages	*Limitations*
1. Validity		
a. Curricular	Careful selection by competent persons after experimentation. Fit typical situation admirably.	Inflexible. Too general in scope to meet fully local requirements, especially in unusual situations.
b. Statistical	With best tests, high.	Criteria often defective. Size of coefficients largely dependent upon group tested.
2. Reliability	With best tests, very high; usually above .90, often above .95. Usually fully objective.	No guarantee of validity. Depends on group tested.
3. Usability		
a. Ease of Administration	Definite procedure, time limits, etc. Economy of time.	Manuals require study, often inadequate.
b. Ease of Scoring	Definite rules, keys, etc. Largely routine.	Take time, monotonous.
c. Ease of Interpretation	Better tests have adequate norms. Useful basis of comparison. Equivalent forms.	Norms often confused with standards. Some norms defective. Norms for various types of schools and levels of ability are usually lacking.
Summary Main points, Pro and Con	Convenience, Comparability. Equivalent forms.	Inflexibility.

termining the present status of the class or school, the best choice will usually be a test battery, such as the Stanford,[9] Metropolitan,[9] Modern School,[10] Public School,[11] or Progressive Achievement

---

[9] Published by World Book Company, Yonkers, New York.

[10] Published by Bureau of Publications, Teachers College, Columbia University New York City.

[11] Published by Public School Publishing Company, Bloomington, Illinois.

AND NONSTANDARDIZED TESTS OF ACHIEVEMENT

NONSTANDARDIZED			
TRADITIONAL		NEW-TYPE	
Advantages	*Limitations*	*Advantages*	*Limitations*
Useful for English, advanced classes; afford language training. Encourage sound study habits.	Limited sampling. Bluffing is possible. Mix language factor in all scores.	Extensive sampling of subject matter. Flexible in use. Discourage bluffing. Easier to prevent and to detect cheating.	Narrow sampling of functions tested, largely memory. Negative learning possible. Piecemeal study encouraged.
	Usually not known.	Compares favorably with standard tests.	Adequate criteria usually lacking.
Inexperienced teachers often do better than with new types.	Average is low; usually .60 to .65.	Compares favorably with standard tests.	No guarantee of validity.
	Subjective scoring.	Objective.	
Easy to prepare. Easy to give.	Lack of uniformity.	Directions rather uniform. Time requirements are well worked out. Economy of time.	Difficult to prepare.
	Slow and uncertain.	Definite rules, keys, etc. Largely routine.	Take time. Monotonous.
	No norms. Meaning doubtful.	Local norms can be derived.	No norms available at beginning.
Useful for part of most tests, and in a few special fields.	Limited sampling. Subjective scoring. Time consuming.	Extensive sampling. Objective scoring. Flexibility.	Narrow sampling of functions tested. Preparation requires skill.

Tests.[12] Any of these tests will give a fairly adequate measure of the present status and the general fitness of the pupils for future work.

For a general survey of the intellectual status of the class or school, a good group test of general intelligence will suffice, although as a rule an average of two is better than one alone. In any measurement of intelligence involving group tests, especially if only one test is used, it is desirable to have retested with an individual intelligence test, such as the Stanford-Binet, the following pupils: those who test very low, say below an IQ of 80; those who test very high, say above an IQ of 130; or those whose scores are considerably out of line with the judgment of the teacher. The Revised Stanford-Binet appears to be particularly trustworthy at the low IQ levels. The distinctive advantage of the individual intelligence test is the opportunity afforded for the examiner to observe the behavior of the child under standardized conditions. As a diagnostic instrument such a test is likely to be much superior to the group test. Any pupils who have language difficulty should be tested with a performance test.

---

[12] Published by California Test Bureau, Los Angeles, California.

A reasonably complete testing program will require, as a rule, the use of general intelligence tests along with achievement tests. Because of the relative constancy of the IQ it is unnecessary to administer intelligence tests each year. The mental level of most pupils can be predicted closely enough from intelligence tests periodically scheduled to permit ordinary comparisons with achievement. Page 194 outlines intelligence testing programs adapted to various types of school organization. At times, aptitude tests in specific fields, rating scales, check lists, personal interviews, and the like will also be required. The particular combination of measuring techniques required in any given situation will depend upon the specific purposes to be served. As a rule, classroom teachers will find a larger place for nonstandardized, teacher-made tests in the solution of instructional problems than will school administrators in the solution of administrative problems. The reverse condition will tend to be true for standardized tests. Table 10 is a sort of "balance sheet" which briefly summarizes some of the chief advantages and limitations of various types of achievement tests. It is evident that there is a legitimate place for all kinds of tests, but no one test is equally good for all purposes.

What is the best procedure? Regardless of the purpose of the testing program or who makes the selection of tests, it is important that a systematic, businesslike procedure be employed. Users of standard tests will find the information contained in *The Mental Measurements Yearbooks*[13] of great value. The comprehensive character of the tests reviewed in this publication is indicated by the

---

[13] Prepared by Oscar K. Buros. The first of the series was published by the Rutgers University Press in 1938. Later editions published by the author.

TABLE 11

OTIS SCALE FOR RATING STANDARD TESTS[14]

SCALE FOR RATING TESTS	NAMES OF TESTS				
Manual (5)					
Validity (15)					
Reliability (10)					
Reputation (5)					
Ease of Administration (Total 15) (a) Preparation (4)					
(b) Time limits (4)					
(c) Explanation needed (3)					
(d) Alternative forms (4)					
Ease of Scoring (Total 15) (a) Objectivity (10)					
(b) Time required (3)					
(c) Simplicity (2)					
Ease of Interpretation (Total 15) (a) Norms (5)					
(b) Directions for interpreting (4)					
(c) Class record (1)					
(d) Application of results (5)					
Convenient Packages (5)					
Typography and Makeup (5)					
Test Service (10)					
Total (100)					

---

[14] Published by World Book Company.

"Table of Contents" of *The Nineteen Forty Mental Measurements Yearbook*[15] as follows:

	PAGE
ACHIEVEMENT BATTERIES	19
CHARACTER AND PERSONALITY	49
ENGLISH	100
Grammar and Usage	100
Literature	125
Speech	135
Spelling	136
Vocabulary	140
FINE ARTS	143
Art	143
Music	150
FOREIGN LANGUAGES	157
French	158
German	179
Italian	182
Latin	182
Spanish	191
INTELLIGENCE	198
MATHEMATICS	268
Algebra	272
Arithmetic	281
Geometry	302
Trigonometry	312
MISCELLANEOUS	314
Agriculture	314
Business Education	314
Computational and Scoring Devices	317
Education	319
Health	320
Home Economics	324
Industrial Arts	329
Record and Report Forms	333
Religious Education	333
Safety Education	333
Sensory-Motor	335
READING	336
SCIENCE	380
Biology	380
Chemistry	385
General Science	390
Physics	396
SOCIAL STUDIES	404
Economics	413
Geography	413
History	415
VOCATIONS	428

As an illustration of the type of evaluations in this volume, the following excerpts from comments on the Traxler High School Reading Tests are given.[16]

Alvin C. Eurich, Professor of Education and now Vice-President of Stanford University, stresses reliability and is not particularly complimentary:

> Clearly the understanding and comprehension sections of this test are not as reliable as available reading tests for high school students, such as the *Iowa Silent Reading Test* and the *Nelson-Denny Reading Test*. The rate section, mainly because it is longer than other similar tests, is slightly more reliable. On the whole, the Traxler test does not provide a better instrument for measuring reading ability than those already available; in fact, it is not as good.

Constance M. McCullough, Assistant Professor of Education, Western Reserve University and an authority on reading, renders a more favorable verdict:

---

[15] Oscar K. Buros (Editor), *The Nineteen Forty Mental Measurements Yearbook*, page xxiii. Highland Park, New Jersey: Mental Measurements Yearbook, 1941.

[16] *Ibid.*, pages 370–373.

> However, while Traxler has limited the scope of his test to a few rather specific reading abilities functioning in two special types of material, he has provided a diagnostic measure more thorough and explicit than most of the tests in this field can claim to be. His contribution is a reminder that reading abilities are many and varying, and that a 40-minute test which attempts to touch upon more than two or three abilities through a given type of material forfeits reliability in its parts and its utility in the schools. The test users need to recognize this fact and to select tests according to their appropriateness to specific situations. If they wish to survey a broad range of reading abilities with varied materials, they must be ready to devote more than an hour's time to testing.

C. Gilbert Wrenn, Professor of Educational Psychology, University of Minnesota, ends his comment with this confident tone:

> This test will be useful in senior high schools, particularly to those educators who have found the Traxler test for grades 7 to 10 helpful. Its value over other published tests will be more apparent when further standardization data are available. It is a carefully made test of conventional form which adds but little to our knowledge of reading-test techniques.

The *Yearbook* also includes quotations from two rather favorable reviews of the test published in educational journals during the preceding year.

In the choice of standard tests it is always wise to have available for careful examination both the test blanks and the test manuals of all tests being considered.[17] To assist in making the necessary examinations and comparisons, the use of a rating scale will be found helpful. The first one published, and still one of the best, is that prepared by Otis, reproduced in Table 11. A more analytical scale for evaluating achievement tests is that of Cole and von Borgersrode, given in Table 12.

Table 12

COLE-VON BORGERSRODE SCALE FOR RATING STANDARDIZED TESTS[18]

I. Preliminary Information

1. Exact name of test.
2. Name and position of author.
3. Name of publisher and nearest address.

---

[17] Most county and city school systems will find it desirable to have available for such purposes complete sample sets of the more important tests published. The following companies issue convenient sets of such materials:

Bureau of Publications, Teachers College, Columbia University, New York City
Cooperative Test Service, New York City
Educational Test Bureau, Minneapolis, Minnesota
Public School Publishing Company, Bloomington, Illinois
World Book Company, Yonkers, New York

[18] Robert D. Cole and Fred von Borgersrode, "A Scale for Rating Standardized Tests," *School of Education Record of the University of North Dakota,* 14: 11–15, October, 1928.

TABLE 12 (CONTINUED)

COLE-VON BORGERSRODE SCALE FOR RATING STANDARDIZED TESTS

I. Preliminary Information (*Cont.*)
  4. Cost.
  5. Date of copyright.
  6. Purpose of test.

II. Validity (25)

  A. Curricular (15)
    1. Exact field or range of educational functions which test measures?
    2. Ages and grades for which intended?
    3. Criteria with which material was correlated?
    4. Do questions parallel good teaching procedures?
    5. How wide is sampling of important topics?
    6. What is the social utility of questions?
    7. Is test claimed to be diagnostic? (If so, proof and see VI, 5, c, below).

  B. Statistical (10)
    1. Correlated against what outside criteria?
    2. Size of coefficient of correlation?
    3. Size and representativeness of sampling?
    4. Proof of validity of items (such as statements as to experimental tryout of items individually to determine that no large percentage is failed or passed by all pupils and that the items show a consistent increase of percentages of successes with successive age or grade levels.)

III. Reliability (25)

  A. Most important items.
    1. Correlated with what?
    2. Size and representativeness of sampling?
    3. Reliability coefficient.
    4. The means of the distributions.
    5. The standard deviations of the distributions.
    6. If some other measure than the above three is given to prove reliability, what is it?
    7. Intercorrelations.

  B. Less important but desirable.
    1. Order of giving various forms of test.
    2. Is test reliable enough statistically for individual measurement, or can it be used only for groups?
    3. Evenness of scaling (see II, B, 4).
    4. Are pupils accustomed to this type of test?

IV. Ease of Administration (15)

  1. Manual of Directions (3)
    a. How complete and simple is the manual?
    b. Does manual control test conditions well?
    c. Typographic makeup.

  2. Simplicity of administration (8)
    a. Amount of explanation needed for pupils by examiner?
    b. Are directions to pupils clear, detailed, comprehensive?
    c. Is arrangement of test convenient for pupils?
    d. Are samples and "fore-exercises" given when needed?

TABLE 12 (CONTINUED)

COLE-VON BORGERSRODE SCALE FOR RATING STANDARDIZED TESTS

IV. Ease of Administration (15) (*Cont.*)
  3. Alternate forms (3)
    a. Number.
    b. Evidence of reliability.
    c. Evidence of equivalency.
  4. Time needed for giving.

V. Ease of Scoring (10)
  1. Degree of objectivity—purely objective or some judgment on part of examiner?
  2. Are adequate directions given—clear, equal to all emergencies?
  3. Is scoring key adjusted to size of test?
  4. Time needed to score one test.
  5. Simplicity of procedure.
    a. Number of processes needed to get final score?

VI. Ease of Interpretation (20)
  1. Norms (6)
    a. Kind—age, grade, percentile, etc.
    b. Derivation—size and representativeness of sampling.
    c. Tentative, arbitrary, or experimental?
    d. For separate parts?
    e. How expressed?
  2. Is class record provided?
  3. Are there provisions for graphing results?
  4. Is interpretation of raw scores easy or hard?
  5. Application of results (10)
    a. Are directions or suggestions given for application of results to benefit teaching or administration?
    b. Are tests survey or diagnostic?
    c. If diagnostic—
      (1) Proof of diagnostic value?
      (2) What principle or principles underlie construction?
      (3) How many different skills, abilities, or aspects of the subject are analyzed or measured?
      (4) Does the analysis of total subjects into unit abilities follow teaching practices of needs?
      (5) Is the diagnosis individual or class—proof?
      (6) Does the test demand tabulations of individual pupils' errors to secure a diagnosis?
      (7) Is a remedial program provided or suggested?

VII. Miscellaneous (5)
  1. Typography and makeup.
    a. Arrangement of printed matter.
    b. Legibility of type.
    c. Quality of paper.
    d. Are test blanks free from distractions, norms, directions to examiner, etc.?
  2. Is the time required for giving as small as is consistent with reliable measurement?
  3. Is the cost in keeping with the amount, scope, and reliability of the results yielded?
  4. Is good test service provided by the publisher?
  5. Kind of new-type questions used?

The use of these scales not only directs attention to significant points but also gives some idea of the relative weight of the various items. However, in the Cole-von Borgersrode scale, note that the first five items under VI, 5, c, definitely refer to curricular validity, rather than to ease of interpretation, which is one of the principal factors in the usability of a test. Furthermore, in the author's opinion, Cole and von Borgersrode assign too much weight to reliability, and both scales assign too little weight to validity, the most important quality of any measuring instrument. Also, the relative weight assigned by both these scales to what may be termed *usability* seems somewhat heavy. The author would suggest a slight revision in weights and a regrouping of sections IV, V, VI, and VII under the heading *usability*. The major divisions and subdivisions, with revised weightings, would then be as follows:

Division		*Points*
I. Preliminary Information		
II. Validity		50
A. Curricular	30	
B. Statistical	20	
III. Reliability		20
A. More important items	15	
B. Less important items	5	
IV. Usability		30
A. Ease of administration	10	
B. Ease of scoring	5	
C. Ease of interpretation	10	
D. Miscellaneous	5	
Total		100

A desirable procedure is to have a group of at least three competent people, each independent of the others, look over all the tests being considered, the manuals accompanying them, and any evaluations available. Each judge first compares the tests with respect to validity, and records the judgment in points before considering anything else. Then he goes on to reliability and makes a similar judgment on each test. Finally, he does the same for usability. This method will tend to produce greater agreement among the judges regarding the *relative ranks* of the tests on the criteria individually. After all, the total point score allowed a test is much less important than the rating on the divisions separately.

Emphasizing the close relationship between teaching and testing, Brownell has suggested the following criteria[19] for evaluating tests:

---

[19] William A. Brownell, "Some Neglected Criteria for Evaluating Classroom Tests," *National Elementary Principal*, 16: 485–492, July, 1937.

1. Does the test elicit from the pupils the desired types of mental processes?
2. Does the test enable the teacher to observe and analyze the thought processes which lie back of the pupils' answers?
3. Does the test encourage the development of desirable study habits?
4. Does the test lead to improved instructional practice?
5. Does the test foster wholesome relationships between teacher and pupils?

In selecting a test for a given purpose, the grade level on which it is to be used must be given consideration. Test publishers often suggest a considerable grade range in which the test may be used. But both test authors and publishers tend to be too optimistic concerning the range of usefulness of their tests. For example, an intelligence test that is supposed to be suitable for grades three to eight may be found to be too difficult for the third grade and too easy for the eighth. It will doubtless be recalled from an earlier discussion that it has usually been found that a test has optimum discrimination for a group whose average score is approximately 50 per cent of the maximum score possible on the test. It is easy to find this point for a test and consequently to select tests of optimum value for each grade. For example, the Terman Group Test of Mental Ability has a maximum score of 220 points. Fifty per cent of this is 110 points. Now look in the table of norms for the mental age equivalent of 110. This is found to be 15 years and 3 months, or about the age of a typical pupil in the ninth grade. The Terman test will then be expected to discriminate best in the ninth grade. It must be remembered, however, that the discriminating function of diagnostic and certain other specific tests is usually relatively unimportant.

C. Administering the Test

The next step in the testing program is the administering of the test. Traxler[20] suggests fourteen practical rules for administering a testing program. In order to insure that this is properly done, three questions must be answered:

1. When should the tests be administered?
2. Who should administer the tests?
3. What is the correct procedure to follow?

Each of these questions deserves careful consideration.

When should the tests be administered? As problems concerning the use of intelligence tests differ somewhat from those concerning the use of achievement tests alone, it is better to consider the two separately. When should intelligence tests be administered?

---

[20] Arthur E. Traxler, *op. cit.*, pages 261–263.

There is general agreement that it is not necessary to give the same pupils intelligence tests every year, but there is also agreement that possible fluctuations on group tests are great enough to warrant giving such tests more than once. The fluctuations are likely to be most serious from the low to intermediate grades.[21] A reasonable plan employed by many school systems is to give intelligence tests at transitional points in the pupil's school history. As Stoddard suggests: "Intelligence is analogous to health; any estimate of it should be rechecked close to the making of an important decision." [22] Procedure would therefore vary according to the school organization. A suggested minimum program is as follows:

Type of Organization	*Grades to Give Intelligence Tests*
Six-six plan	First and sixth or seventh
Seven-four plan	First and seventh or eighth
Eight-four plan	First and eighth or ninth
Six-three-three plan	First, sixth or seventh, and ninth or tenth

If possible, it would be well to add to this minimum program a test at about the fourth grade and one at the end of the high-school course.

There is some disagreement regarding the best time of year in which to give the intelligence tests. Of course, if the tests are to have maximum value, their results must be made available at the very beginning of these transitional periods. This means they should be given early in the first grade if the pupils have had no previous kindergarten experience. Since Updegraff [23] found that for preschool children the reliability of the test is increased by postponing testing until two weeks after entrance to school, it may be well to avoid giving the test till the second or third week of school in the lower grades. The later tests can be given either at the beginning of the transitional year or at the close of the year preceding. There is a tendency to have tests for college entrance administered in the high schools near the close of the senior year. This is obviously necessary if such tests are to be used in counseling these seniors regarding the possibility of continuing their education. There will usually be a few pupils who will transfer into the system and who have not had intelligence tests, and others in the system about whom teachers may feel serious doubt regarding the validity of the existing record. It is a good rule to retest any pupil whose IQ varies more than 10 points from that on an earlier test. Such cases, which will usually be few in number, should be tested when the need arises.

---

[21] *Cf.* Mildred M. Allen, "Relationship between the Indices of Intelligence Derived from Kuhlmann-Anderson Intelligence Tests for Grade I and the Same Tests for Grade IV," *Journal of Educational Psychology,* 36: 252–256, April, 1945.

[22] George D. Stoddard, *The Meaning of Intelligence,* page 94. New York: The Macmillan Company, 1943.

[23] Ruth Updegraff, "The Determination of a Reliable Intelligence Quotient for the Young Child," *Pedagogical Seminary and Journal of Genetic Psychology,* 41: 152–166, September, 1932.

The frequency with which achievement tests should be used will depend primarily upon the purpose they are to serve. Most purposes, however, will require at least two series of tests administered at intervals of a semester or a year. Most achievement tests have norms for the middle and the end of the year, but often for no other time. When tests are given at these periods, comparisons with norms are easiest. There is also the fact that many studies have shown a considerable decline in knowledge at the end of the summer vacation. This would seem to favor giving the tests at the end of the school year, when the pupils' status is more normal. A comparison between the records made by pupils at the end of each of two successive years is usually more trustworthy than that between the beginning and end of one year.

There are some advantages in having the tests administered in the fall. Almost always some pupils will enter the school for the first time and their status can best be determined by administering tests to all the pupils. The teachers will then have the entire school year in which to remedy any deficiencies revealed. Fall testing also avoids the undesirable practice of cramming. If too much emphasis is placed on "improvement," shown during the year, however, pupils may be tempted not to do their best on the first series of tests. This would not be the case if progress is measured between two series of tests administered at the end of the preceding year and at the end of the current school year.

This practice will also make it possible to have the information serve several purposes. It can be used partially as a basis for determining promotion from the grade, for educational guidance, and for proper sectioning in the next grade. There seems also no good reason why an analysis of the errors revealed cannot serve equally well as a basis for remedial teaching in the succeeding grade as if the new teacher had given the test at the beginning of the year. Of course, in some instances there might be considerable value in repeating the test at the beginning of the year in order to determine the deteriorating effects of the summer vacation, apart from the better-established weaknesses which were present when the vacation started. Moreover, the analysis of errors is more trustworthy when based upon two samplings of performance than upon one.

Who should administer the test? It goes without saying that only competent persons should administer standardized tests. It is not always an easy matter to tell who is really competent, however. In the case of individual tests of the Stanford-Binet type, this requirement means that only persons who have had specific instruction in college classes should attempt to administer them. There should be at least one person in every school who is qualified to give such tests. When tests are used for purposes of research, or when they are used to compare one grade, class, or school with others, they should usually be given by one person, or a small group of specially trained examiners. But in the ordinary testing program, employing group intelligence tests and achievement tests, the regular classroom teachers should usually administer the tests. Most of them will welcome an opportunity to do so. At the present time there seems no good reason for selecting a test whose administration is so difficult as to be beyond the mastery of average teachers in the public schools. The point of view of McCall seems eminently sound:[24]

> Many years ago certain specialists sought to secure a monopoly of the privilege of using standard tests by trying to persuade educators to regard the tests as possessing certain mystic properties. A few of us with Promethean tendencies set about taking these sacred cows away from the gods and giving them to mortals. Can teachers be entrusted with tests? If not, then teachers ought not to be trusted with 90 per cent of their present functions. We now entrust them with the far more difficult task of teaching reading, creating concepts and building ideals. Let us not strain at a gnat when we have swallowed fifty elephants.

But it is well not to take the competency of the examiners for granted. One of the best plans is to get the group of examiners together and demonstrate the administration of the tests to be used. One way to do this is to give a demonstration with a regular class and to follow this by a discussion with the examiners of the procedure they have seen. Another way is to administer the test to the examiners themselves, using somewhat shortened time allowances. This should be followed by a full discussion of the procedure involved. It is usually well to suggest that after each examiner has studied the manual he try the procedure on some other person, such as a member of the family; or two teachers may try it out on each other. If questions then arise, they can be settled by a conference with the person in general charge of the program before the examiner goes before his group actually to administer the test. It has been found that, if such measures are taken, the regular classroom teachers can obtain practically the same results as can be obtained by special examiners when group tests are employed.

---

[24] W. A. McCall, in *The Test Newsletter,* published by Bureau of Publications, Teachers College, Columbia University, December, 1936.

What procedure should be followed? Although the procedure of administering group intelligence tests and achievement tests is not beyond the mastery of classroom teachers and school administrators, some difficulties may arise. In fact Ligon[25] argues that good group testing is more difficult than individual testing. In the first place, the conditions for the test must be favorable. It is usually best to have the tests given in the familiar environment of the pupils' own classrooms. Especially is this true of younger children. It is well always to have the tests given at regular class time without permitting them to run over into lunch hour or play time. For the same reason it is desirable not to have tests just before or just after an important event, such as a holiday, a school party, or an athletic contest. Precautions should be taken to avoid all unnecessary distractions and interruptions during the progress of the test. It is a good plan to hang on the outside of the classroom door a card which reads: *Tests Going On. Please Do Not Disturb.* Pupils should be instructed to remove everything from the tops of their desks except one or two well-sharpened pencils and an eraser. The examiner should also have ready a few extra pencils in case of an emergency. All these things must be looked after in order to insure favorable working conditions for the test.

As a rule, anyone can administer a group test successfully who meets three requirements. The first of these is the ability to read well. Good silent reading is required for the mastery of the directions printed in the manual which accompanies the test. Good oral reading ability is required, for the directions to the pupils should be *read,* not recited from memory. To undertake to give the test from memory is to run a serious risk of leaving out some important word or phrase or of paraphrasing the directions in such a way as to change their meaning. But the examiner should be so familiar with the manual that he can read the directions with his eyes off the page a good part of the time. The directions should be read with proper emphasis in a clear voice just loud enough to be heard throughout the room. The aim should be to make the meaning understood without arousing anxiety or excitement.

The second requirement for administering a test is the ability to keep time accurately. To accomplish this a stop watch is desirable, or at any rate a watch with a second hand. The aim should be to keep the time to a second. On most tests the signal to start is, "Ready, go!" or "Ready, begin!" When this signal is given, the examiner should note the *exact* time—hour, minute, and second.

---

[25] Ernest M. Ligon, "The Administration of Group Tests," *Educational and Psychological Measurement,* 2: 387–399, October, 1942.

This should be *immediately recorded,* preferably on a small card or specially prepared blank. The record for Test 1 would look like this:

Test 1	*Hr.*	*Min.*	*Sec.*
Time test began	8	20	15
Time allowed		5	
Time to stop	8	25	15

Experienced examiners know that it is never safe to trust one's memory to keep the time. A written record must be made.

The third requirement for administering a test is the ability to follow directions accurately. The manual should be followed verbatim. No deviation whatsoever is permissible. To add anything to, or to modify the directions in any way, means that it is no longer a standardized test. Boynton [26] gives some interesting illustrations of unconscious clues given by inexperienced examiners using the Stanford-Binet. One examiner, for example, when asking the meaning of the word "tap" in the vocabulary test began to tap on the table, and when he came to the word "eyelash" he looked the child straight in the eye and batted his eyes rapidly. The norms are made on the assumption that a prescribed formula is to be used. As a part of the preliminary instructions pupils are almost always told not to ask any questions after the test starts. Occasionally a pupil forgets this instruction and holds up his hand for a question. The examiner should walk over to him and, if it is a reading test or an intelligence test whose purpose is following directions, should say in a quiet voice, "Read it carefully and do just what it says." If it is an ordinary achievement test and the pupil is concerned about where to put his answer or some other point of mechanics that does not involve the answer to a question in the test or modify the directions already given, it is permissible to set the pupil at ease without causing disturbance. Kelley suggests this principle in handling the child who is in trouble: "The examiner should be free to say or do anything that does not disturb or delay pupils at work, that does not help the individual child in the thing in which he is being tested, and that does set him to work again after some foolish or trivial issue has troubled him." [27] Examples of permissible statements are: "Yes, you may change your response if you decide it is wrong," "Just work on the side of the sheet, you do not need scratch paper," "When you have finished the first column go right on to the next one," "No, you must not go back to a test you have passed," and the like. But if the pupil asks the meaning or spelling of a word, or how to answer a test item, the examiner should say quietly: "I cannot tell you. Go on to the next one." *In case of doubt, the examiner should err on the side of saying nothing.* While the test is in progress the examiner must be alert constantly to see that the pupils neither help nor hinder each other nor are distracted by external factors. Ligon[28] indicates the following requirements of good group testing: "That all the subjects understand the instructions, that they all work throughout the assigned time at their optimum level of achievement, that they are in no way helped, hindered, or distracted by one another, that they do not quit trying or omit any section of the test, that examiners give instructions adequately and in a stimulating, effective tone of voice—not a dull bored monotone—and that proctors are observing every movement of the group, stimulating lagging souls, inhibiting wandering eyes, and detecting failure to follow instructions." A test is more than a measuring device; it presents a standardized situation in which to observe pupil behavior. Any occurrence observed during the progress of the test that may throw light upon the interpretation of the results should be carefully recorded.

---

[26] Paul L. Boynton, *Intelligence, Its Manifestation and Measurement,* pages 276–277. New York: D. Appleton-Century Company, 1933.

[27] Truman Lee Kelley, *Interpretation of Educational Measurements,* page 46. Yonkers: World Book Company, 1927.

D. Scoring the Tests

It is desirable to have the tests scored as quickly as possible and with the highest possible degree of accuracy. As a rule, then, that system is best which accomplishes these objectives with the minimum expenditure of money, time, and energy. There are two questions involved:

1. Who should score the tests?
2. What technique should be used?

Who should score the tests? In actual practice, standard tests are scored by a variety of persons. Sometimes, especially in larger systems, the work is done by a clerical staff at a central bureau, or by the use of scoring machines; sometimes it is done by advanced students under supervision; at other times the scoring is done by administrative officials; but the most common method seems to be to have the work done by the regular teachers. Lee and Segel[29] found that informal tests were corrected by 92 per cent of the teachers and standard tests by 85 per cent. Except in the larger systems where there is a bureau of research equipped with special facilities, the scoring is probably best done by the classroom teachers. In that way not only can the work be done promptly, but the teachers can probably learn something of value about the types of errors made on the achievement tests. But it is important to get the scoring done without producing an unfavorable attitude toward it on the part of the teachers. Some schools have found it very satisfactory to dismiss classes at noon when the testing is in progress, so that the teachers can devote the afternoon to the work of scoring. This would seem an effective way of emphasizing the important fact that teaching and testing are processes that are intimately related.

---

[28] Ernest M. Ligon, *op. cit.*, page 387.

[29] J. Murray Lee and David Segel, *Testing Practices of High-School Teachers*, page 17. United States Office of Education Bulletin, No. 9, 1936.

What techniques should be used? Every reasonable precaution should be taken to assure a high degree of accuracy in scoring. It must not be assumed that merely because the directions are clear, the key complete, and the process entirely objective, perfect protection against errors is thereby afforded. Studies by Morrison,[30] Pintner,[31] Madsen,[32] and Dearborn and Smith[33] give abundant evidence to contradict this assumption. The conclusions of Dearborn and Smith are quoted[34] as representative:

> 1. Many errors in scoring were present.
>
> 2. There was enough constant error to affect the entire group, and to make conclusions based upon individual mental ages computed from unchecked test scores open to doubt.
>
> 3. There was persistent underscoring.
>
> 4. Objectivity of scoring is not sufficient insurance against error.
>
> 5. There is no sure way to prevent error in scoring, though frequent and careful *instruction* of scorers is very helpful and is probably essential to accuracy in scoring.
>
> 6. A check on scoring, preferably by someone other than the original scorer, is essential to accuracy in scoring.

These studies reveal two distinct types of errors in scoring, *constant errors* and *variable errors.* A common example of the former type is misunderstanding the scoring directions; for instance, by counting omissions the same as errors, in using the scoring or correction formula. Such errors are especially serious, because there is nc possibility of their offsetting each other according to any so-called "law of averages." Variable errors, on the other hand, sometimes tend to make the score too high and at other times too low. While such errors may do serious harm to individual pupils, they tend to cancel each other in group measures such as averages. Examples of variable errors are errors resulting from carelessness, errors in counting the scores, errors in entering the scores on the front of the test booklet or on the record sheet, and errors in adding up the total score. Some of the most serious errors found are not in marking the paper at all but in counting and in addition.

---

[30] J. Cayce Morrison, "Teachers' Errors in Scoring the Illinois Intelligence Scale," *Educational Research Bulletin,* 4: 55–58, February 4, 1925.

[31] Rudolph Pintner, "Accuracy in Scoring Group Intelligence Tests," *Journal of Educational Psychology,* 17: 470–475, October, 1926.

[32] I. N. Madsen, "Participation in Testing Programs by the Classroom Teacher," *Educational Administration and Supervision,* 15: 117–126, February, 1929.

[33] Walter F. Dearborn and C. Wilson Smith, "The Results of Rescoring Five Hundred Thirty Dearborn Tests," *Journal of Educational Psychology,* 20: 177–183, March, 1929.

[34] *Ibid.,* pages 182–183.

Clearly, then, accuracy in scoring cannot be taken for granted. What is to be done about it? The first thing is to prevent the occurrence of errors whenever possible. The scorers must be *taught* how to score the papers and not merely *told* how to do it. They should be given an opportunity to study the manual and the scoring keys. Whenever possible, an actual demonstration of scoring should follow. It is a good idea, also, to check carefully the first few papers marked by beginners to detect errors at the outset. This procedure should reveal any constant errors and the principal types of variable errors. It is always desirable to have each page or part of the test scored through all the papers in a set before going on to the second page or part of the test. If the scorers work in groups, as is usually desirable, each one can specialize in marking one part of the test, and pass the test when scored to the next scorer, who is specializing in marking the next part of the test. This procedure will reduce the risk of error and at the same time will increase the speed of scoring. It is usually an especially poor technique to have one person read the answers while the scorers mark the papers. This is slow, because the slowest scorer sets the pace. It also increases the risk of error, owing to the possibilities of losing the place or of failure to hear correctly. Colored pencils are desirable. Inexperienced scorers should mark each item in the test being scored in some uniform manner, such as + for correct, — for incorrect, and 0 for omitted items. Experienced scorers will save time by marking only the incorrect and omitted items. It is, of course, unnecessary to mark the items below the last one the pupil attempts. But it is well to draw a horizontal line across the test under the last item attempted. Figure 4 illustrates the scoring of Test 3 of the Terman Group Test of Mental Ability, Form A.

The writer has found that the simple device of keeping a written record of who marks, checks, transcribes, or totals each part of the test reduces the likelihood of error. If the scoring is organized systematically, it is a simple matter to keep such a record on a mimeographed sheet attached to each package of tests when scored.

Figure 5 gives a sample scoring record for the Modern School Achievement Tests.

But in spite of these preventive measures, certain errors are likely to occur. The safest plan, therefore, is to have each set of papers marked a second time by different scorers, using pencils of a different color. Dunlap [35] found that items most subject to errors in scoring are of the two-response type requiring a scoring formula and items requiring the underlining of more than one word. If a complete rescoring does not seem practical, a sampling method may be followed. Each fifth or tenth paper, for example, may be selected and carefully rescored, and if only an occasional minor error is found, the whole set may be safely accepted. On the other hand, if frequent or serious errors are found in these sample papers, the entire set should be rescored. In any event it is important to have some person, other than the original scorer, check the totals for each part of the test and for the whole test, all substitutions in the scoring formulas, all transcribing of scores, and all transmuting of point scores into derived scores.[36] It is possible to locate many serious errors by examining closely the profile of each individual pupil on all tests with this form of record. Any score much higher or much lower than the general level is suspicious. Also, when two or more tests are used which purport to measure the same function, any serious discrepancies should be scrutinized, on the supposition that a high positive correlation is to be expected. The standard of absolute accuracy should be accepted by all scorers. The possibilities of serious injustice to individual pupils by errors in scoring should be fully recognized.

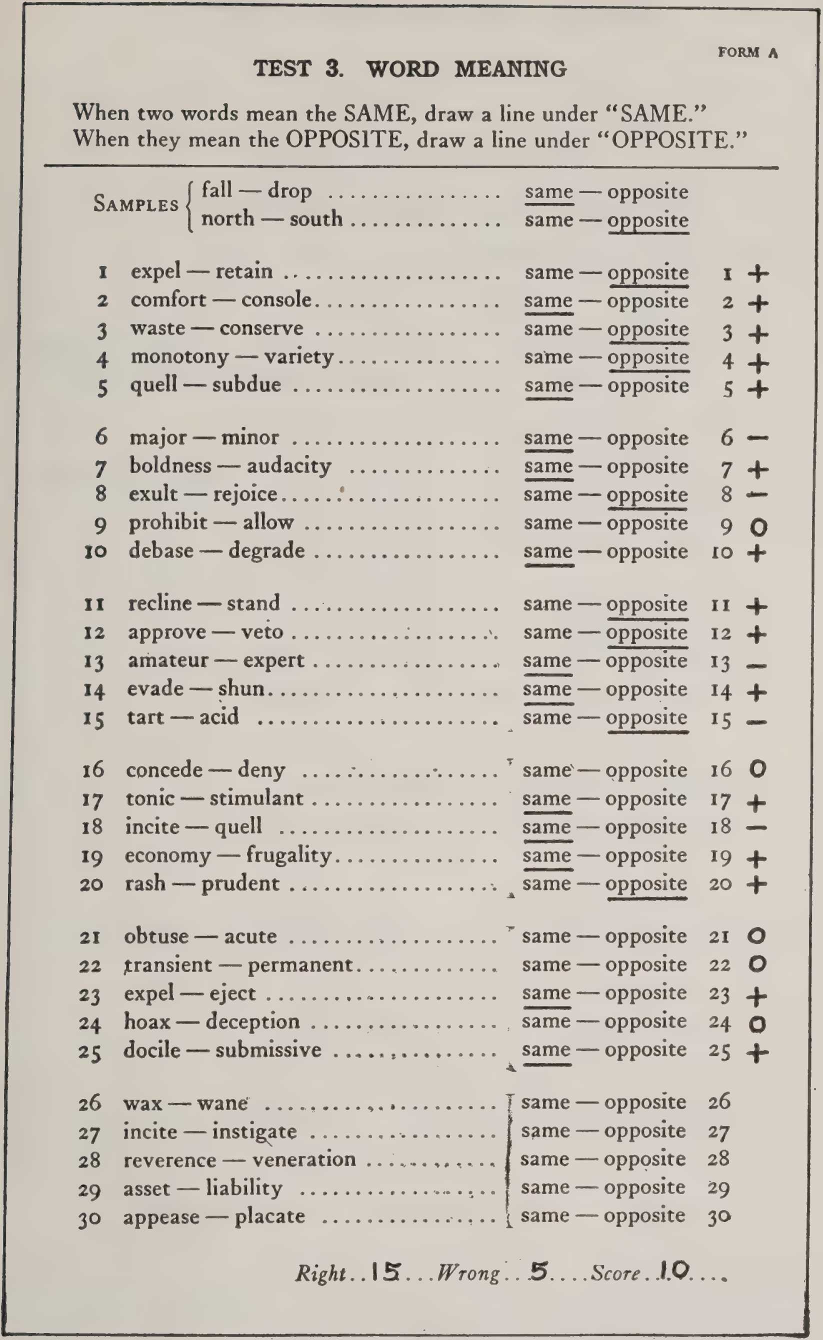
E. Analyzing and Interpreting the Scores

After the tests have been scored and checked, the next step is the analysis and interpretation of the results. Both processes go on together, for analysis is worthless without interpretation and interpretation is impossible without analysis. Analysis is of two main types, statistical and graphical. Before either can be undertaken, however, there is the important preliminary step of classification and tabulation. An analysis of errors appearing in the test papers is usually of major importance to the classroom teacher. The raw scores on standard tests are converted into derived scores by the use of tables of norms before the statistical analysis begins. As the

---

[35] Jack W. Dunlap, "The Relationship Between the Type of Question and Scoring Errors," *Journal of Experimental Education,* 6: 376–379, March, 1938.

[36] Derived scores are obtained from tables of norms. Each point score is expressed in some equivalent unit, such as an age or percentile score. The interpretation of these units is considered in Chapter X



FORM A

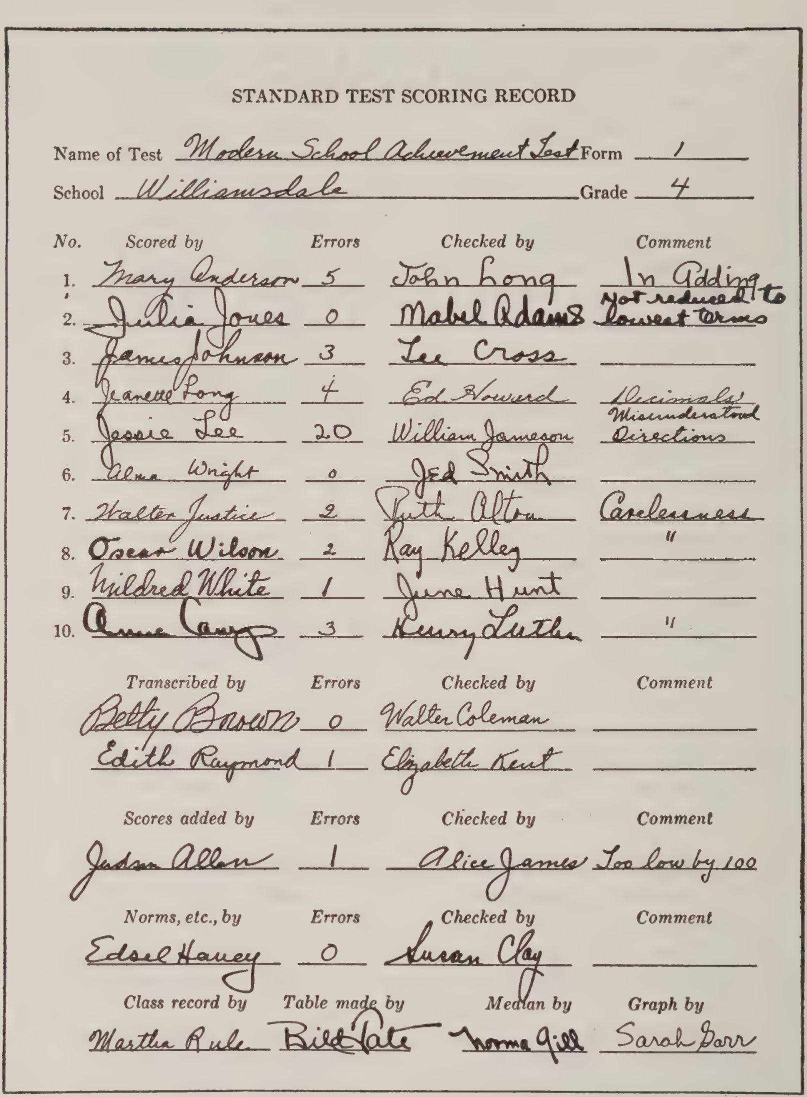
TEST 3. WORD MEANING

When two words mean the SAME, draw a line under "SAME."
When they mean the OPPOSITE, draw a line under "OPPOSITE."

SAMPLES	fall — drop	[✓ same] — opposite		
	north — south	same — [✓ opposite]		
1	expel — retain	same — [✓ opposite]	1	+
2	comfort — console	[✓ same] — opposite	2	+
3	waste — conserve	same — [✓ opposite]	3	+
4	monotony — variety	same — [✓ opposite]	4	+
5	quell — subdue	[✓ same] — opposite	5	+
6	major — minor	[✓ same] — opposite	6	—
7	boldness — audacity	[✓ same] — opposite	7	+
8	exult — rejoice	same — [✓ opposite]	8	—
9	prohibit — allow	same — opposite	9	O
10	debase — degrade	[✓ same] — opposite	10	+
11	recline — stand	same — [✓ opposite]	11	+
12	approve — veto	same — [✓ opposite]	12	+
13	amateur — expert	[✓ same] — opposite	13	—
14	evade — shun	[✓ same] — opposite	14	+
15	tart — acid	same — [✓ opposite]	15	—
16	concede — deny	same — opposite	16	O
17	tonic — stimulant	[✓ same] — opposite	17	+
18	incite — quell	[✓ same] — opposite	18	—
19	economy — frugality	[✓ same] — opposite	19	+
20	rash — prudent	same — [✓ opposite]	20	+
21	obtuse — acute	same — opposite	21	O
22	transient — permanent	same — opposite	22	O
23	expel — eject	[✓ same] — opposite	23	+
24	hoax — deception	same — opposite	24	O
25	docile — submissive	[✓ same] — opposite	25	+
26	wax — wane	same — opposite	26	
27	incite — instigate	same — opposite	27	
28	reverence — veneration	same — opposite	28	
29	asset — liability	same — opposite	29	
30	appease — placate	same — opposite	30	

Right..15...*Wrong*..5....*Score*..10....

Figure 4. An Illustration of the Procedure Followed in Scoring Test 3 of the Terman Group Test of Mental Ability, Form A. (Copyright by World Book Company.)



STANDARD TEST SCORING RECORD

Name of Test Modern School Achievement Test Form 1

School Williamsdale Grade 4

No.	Scored by	Errors	Checked by	Comment
1.	Mary Anderson	5	John Long	In Adding
2.	Julia Jones	0	Mabel Adams	not reduced to lowest terms
3.	James Johnson	3	Lee Cross	
4.	Jeanette Long	4	Ed Howard	Decimals
5.	Jessie Lee	20	William Jameson	Misunderstood Directions
6.	Alma Wright	0	Jed Smith	
7.	Walter Justice	2	Ruth Alton	Carelessness
8.	Oscar Wilson	2	Ray Kelley	"
9.	Mildred White	1	June Hunt	
10.	Anne Carey	3	Henry Luther	"

Transcribed by	Errors	Checked by	Comment
Betty Brown	0	Walter Coleman	
Edith Raymond	1	Elizabeth Kent	

Scores added by	Errors	Checked by	Comment
Judson Allen	1	Alice James	Too low by 100

Norms, etc., by	Errors	Checked by	Comment
Edsel Haney	0	Susan Clay	

Class record by	Table made by	Median by	Graph by
Martha Rule	Bill Tate	Norma Gill	Sarah Dorr

Figure 5. A Sample Standard Test Scoring Record.

next three chapters are concerned with a discussion of the whole problem of analysis and interpretation, only an outline will be given here to indicate the steps involved:

1. Classification and tabulation of scores.
2. Statistical analysis of scores.

3. Graphical analysis and representation.
4. Use of norms and standards.
5. Analysis of errors.

In a complete testing program all five of these steps will receive attention although not always to the same degree. If the primary purpose of the testing program is diagnosis, for example, the fourth step would be relatively unimportant and the fifth step relatively important. The reverse would be true of a program whose main objective is a study of the comparative efficiency of various grades, classes, and schools.

F. Applying the Results

The application of the results is the crux of the whole testing program. Everything that has gone before is really preliminary. Whatever value the tests are to have depends in the last analysis upon the use made of the results.

Just what is to be done, of course, depends upon the purpose of the program. Later chapters will consider in some detail the procedure to be followed for several administrative and instructional problems. It will be sufficient at this point to give some idea of how the procedure will vary with the purpose.

Suppose, for example, that the purpose of the tests is to determine the present classification status of a particular school with the idea of its improvement, and that the test data before the principal make it evident that the situation is far from satisfactory. The question now is, what is to be done about it? Upon the basis of the test scores and other pertinent data, such as the teachers' estimates, health reports, age-grade status, and the like, several pupils are given trial promotions to the next higher grades. A small group of pupils, whose achievement and intelligence scores are well below the central tendency of their respective grades, are organized into an ungraded class and put in charge of a teacher whose outstanding virtues are sympathy, patience, and common sense. Ability groups are also organized in a few grades and classes, with appropriate differentiation in curricula and methods.

Likewise, suppose the primary purpose of the testing program is to determine whether or not the teaching emphasis is correct in the various subjects in the grades, and, when the test results are in, it is apparent that most of the grades are strong in arithmetic and spelling, about normal in reading, and weak in language and the social studies. Now what is to be done here? The principal calls the teachers together and presents the situation in tables and graphs, with suitable comments by way of interpretation. Then follows a regular "council of war." Several committees are appointed to make a special study of the situation and to make recommendations at a meeting to be held a little later. Eventually, after discussion and deliberation, a course of action is decided upon, looking to the improvement of the situation in the weaker subjects.

The procedure will again be somewhat different in essential respects if the primary purpose is diagnosis and remedial work in reading. Here the test results should be analyzed in some detail in each grade. An analysis of the test papers, item by item, is often very revealing. Special effort should be made to locate the specific nature of the reading difficulties. There may be found some general weaknesses, such as the inability to use the index and table of contents in a book, or possibly to locate the central idea in a paragraph. There are usually, in addition, other weaknesses, which appear in certain pupils and not in others. Some of these will not be revealed at all by the usual paper and pencil reading tests, but will require special tools and techniques. After considering these facts, the staff must agree upon a remedial program to be followed during the year.

The essential point in all these cases is that *something is done about the situation revealed by the test scores.* To fail to apply the results in some practical way is to fail in the testing program.

G. Retesting to Determine the Success of the Program

Most testing programs stop with applying the results, if, indeed, they go that far. But an essential step yet remains. After a reasonable time has been allowed for a trial of the remedial measures which were agreed upon in the light of the test data, a checkup should be made to determine the success of this program. Most tests are not sufficiently accurate to reveal progress over a shorter period than one half year. As a rule, a second form of the test or tests used in the beginning should be employed in retesting. If this is not done, it will usually be very difficult to express the results in terms sufficiently comparable to make an accurate measure of progress possible. Of course, not all the gain found can be correctly attributed solely to the remedial program. Some of it is doubtless due to the practice effect or to familiarity with the test itself, part of it to teaching received outside the school, and part of it to natural growth. Often, however, the improvement will be so marked as to indicate beyond a reasonable doubt the effectiveness of the program attempted. At other times the improvement will be disappointingly small. It is then usually wise to modify the remedial program in the light of the results obtained.

The essential point is that the success of the remedial program must not be taken for granted. On the contrary, a definite effort must be made to check upon its effectiveness. To fail to do this is to leave the testing program incomplete. There is no better reason for taking the efficiency of the remedial program on faith than there was for taking the earlier results of teaching on faith.

H. Making Suitable Records and Reports

Certain records and reports are essential to the success of the testing program. But by no means do all these records and reports come chronologically at the end of the program. As a matter of fact, some of these are essential to the last three stages already discussed.

In general, it may be said that four groups have an interest in knowing what the tests show: namely, the pupils, the teachers, the administrative officers, and the parents or public. The nature of the report will naturally vary somewhat with the group to whom it is made, and the nature of the record with the specific function it is to serve. However, regardless of the type of record or its specific function in any particular situation, its general function is always, as has been well stated by Stenquist, "to present test results and related information in such a meaningful way as to *arouse interest and action,* on the part of teachers, principals, supervisors, directors of special divisions, and superintendents."[37]

Report to pupils. The pupils have a right to know their performance on all achievement tests whether standardized or nonstandardized. In many cases it is well to go over the papers with the pupils in order to point out the nature of the errors made. The success of any remedial program will depend upon the pupils' cooperation. Thorndike states the matter succinctly in these words:[38]

> The final justification for every testing regime rests in Mary Jones and John Smith, and it therefore behooves all persons who are making and giving tests to take them into partnership as soon and as completely as is feasible.

It is usually considered dangerous to present the results of intelligence tests to pupils. And there doubtless is more possibility of harm than of good in making known the mental ages and intelligence quotients of individual pupils. Difficulty is most likely to result from scores at the extremes of the distribution. Both pupils and parents can reconcile themselves to low scores on achievement tests, for that can be explained to their satisfaction on the ground that it is the school's fault. But low intelligence test scores seem to reflect directly upon the good name of the family, and this is resented. Only the exceptional pupil or parent has a fine enough philosophy of life to reconcile himself to the realities implied by a low score, and to resolve to make the most of it. There is also danger that the pupils with high test scores will be so inordinately puffed up as to endanger both their social standing with their fellows and their academic standing with their teachers. There are, however, special cases in which information regarding intelligence scores may properly be given. Some examples of these cases will be discussed in later chapters.

---

[37] John L. Stenquist, "The Administration of a Program of Diagnosis and Remedial Instruction," *Thirty-Fourth Yearbook of the National Society for the Study of Education,* page 518. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1935.

[38] Edward L. Thorndike, "Tests and Their Uses," *Teachers College Record,* 26: 93–94, October, 1924.

Records and reports to teachers. The classroom teachers need to have several kinds of records and information, most of which they can prepare themselves. Each teacher needs, first of all, a complete test record sheet for all his pupils. This sheet gives the test record of all members of the class arranged in descending order on the basis of the total score or on the basis of the previous teacher's rating, as is the practice in Baltimore. Most publishers of standard tests include such a record form with each package of twenty-five tests. Some writers recommend an alphabetical order, but the rank order is more useful. Stenquist[39] has described how a testing program in a large city which involves a quarter of a million tests annually is made to function smoothly. "We have found," says Stenquist, "that the effective use of tests is a collective enterprise involving a high degree of cooperation on the part of pupils, teachers, principals, supervisors, directors and superintendents." He also describes the "device that more than all else 'sold' our testing program to teachers, principals and all others concerned."[40] This is simply an analysis chart for each class showing the names, ages, and distribution of scores on each test by entering the identification number of each pupil opposite his score. Duplicate copies are quickly prepared on blueprint machines and sent to the teacher, the principal, the supervisor, and the superintendent.

Most tests which attempt to measure various aspects of a subject, and all test batteries, provide on each test a form for a graphical record of each pupil's performance. Figure 6 shows a sample record for the Metropolitan Achievement Tests, Intermediate Battery. This record enables the teacher to see at a glance not only the pupil's general level, but a picture of his strong and weak points as well. When the results of the later tests have been given and entered on the same sheet in a different color, a clear picture of the pupil's progress is available. Such a record has obvious advantages.

---

[39] John L. Stenquist, "Making Tests Effective," *The Nation's Schools,* 20: 18–21, September, 1937.

[40] John L. Stenquist, "Devices for Testing," *The Nation's Schools,* 20: 30–33, November, 1937.

Diederich suggests a summary report by the teacher, the nature and function of which is described as follows:[41]

> After every important test or examination, whether standardized or home-made, the teacher would do well to prepare a brief report covering the nature of the group which took the test, the nature of the test and how the group was prepared for it, the highest, lowest, and middle scores, and the national norms if they are available. This statement might be mimeographed and one copy put in the folder of each pupil who took the test. On these copies should be typed or written the pupil's score or standing in the test, what this meant, if anything, with relation to the objectives of the course, and some comment as to strengths and weaknesses shown, progress or decline, and possible reasons. Such statements should not take long to prepare, and they would be immediately valuable in counseling. Perhaps no other occasion in the normal processes of school life offers such rich opportunities for helpful counseling. If tests and examinations are worth giving, they are worth recording and interpreting in a form which will enable those responsible for the pupil's education to act intelligently upon them, and to draw sound conclusions from them.

Records and reports for administrators. In a small school the principal will find useful much of the same kind of information for all pupils in the school that the various teachers find useful for their own pupils. In the larger schools and school systems the administrative officers will be mainly interested, not so much in individuals as in the summaries of classes, grades, and schools.

The most important records in the principal's office are the individual records of each pupil in the school. To be most useful these records should be comprehensive, cumulative, and convenient. They should be comprehensive, including not only the test record but other pertinent information regarding the pupil, such as school marks, health record, personality ratings, vocational aptitude and interest data, avocational experience and interests, social background, age and grade progress, and the like. A notable survey of 870 schools by Leonard and Tucker[42] reveals that 83 per cent kept complete records of all pupils permanently. The study also shows that the typical school used six tests, of which intelligence tests were most common. These records should be cumulative; that is, they

---

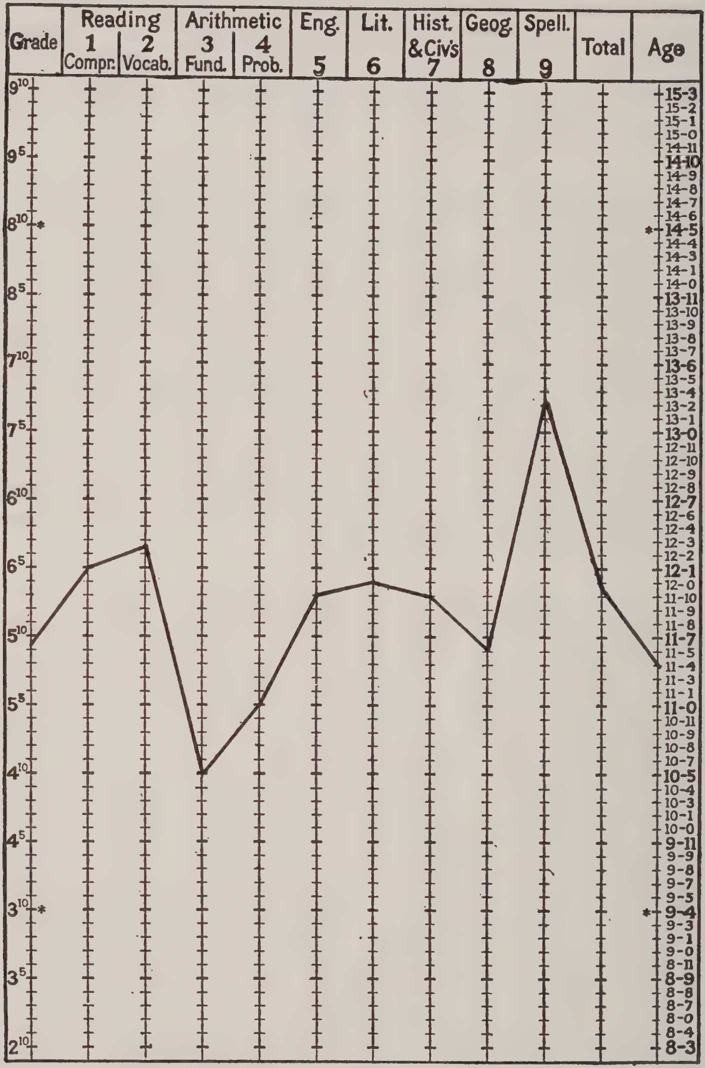
[41] Paul B. Diederich, "Evaluation Records," *Educational Method,* 15: 439, May, 1936.

[42] E. A. Leonard and A. C. Tucker, *The Individual Inventory in Guidance Programs in Secondary Schools,* 60 pages. Washington: U. S. Office of Education, 1941.

Name . Mary . . Smith . . . . *Gr.* 5 . *Age* 11^4 . *Date* May. 15.1940
Teacher Miss . Jones . . . . *School* Lincoln . *City* Lexington

INDIVIDUAL PROFILE CHART

METROPOLITAN ACHIEVEMENT TESTS: INTERMEDIATE BATTERY—COMPLETE



* Values above Grade 8^{10} and below Grade 3^{10} are extrapolated.

The Profile Chart is designed to furnish a graphic picture of the achievement of an individual pupil as given in the table on the front page. The grade equivalents for the test scores, which are needed for the completion of the profile, are obtained from the norms at the end of each test, or they may be obtained from a table of norms based on the local school medians. The Supervisor's Manual should be consulted for further details concerning the interpretation of the Profile Chart and the uses of the test results.

Figure 6. An Educational Profile for a Standardized Achievement Test. (Published by World Book Company, 1933.)

should reveal the pupil's record over a period of years, preferably from the kindergarten through high school. There is a decided advantage in having the complete developmental history of the individual pupil. One study,[43] for example, executed by the author, called attention to the value of the information contained on the ordinary school record card, which gives a "picture of the pupil under varying conditions and stages of development" quite analogous to the biologist's record of the life history of an organism.

Such records must also be convenient to use. For each pupil there should be a card or folder, upon which all data are expressed in comparable units. Figure 7 shows the test data summary from the cumulative guidance record of the Department of Supervision and Curriculum Revision of the N. E. A.[44] There are many advantages in a graphical record, of which type the one published by the Educational Records Bureau is widely used. The test record section is reproduced in Figure 8. Such a graphical record makes it possible to see at a glance the quality, amount, and consistency of progress made. A distinguishing feature of this record is the use of percentiles, so spaced that equal distances vertically represent equal amounts in achievement. Wood, under whose direction the form was devised, makes this significant statement regarding its use: [45]

> The idea that such data should be systematically recorded in comparable and meaningful terms, and that they should be made the subject of continuous and prayerful study and long-time planning on the part of the highest educational officers, as well as of the teachers throughout the whole educational ladder, has been slow to emerge.

Such records, although easy to interpret, are somewhat laborious to prepare. Hangen [46] found that the graphical record required twice as much time to prepare as the numerical record and was somewhat more subject to error. This is perhaps largely responsible for the discovery a few years ago that only about one third of the schools which have membership in the Educational Records Bureau make a graph of all test results.[47] Care must be exercised that such records do not become ends in themselves, and that so much time is not devoted to their preparation that none remains for their

---

[43] Clay Campbell Ross, *The Relation between Grade School Record and High School Achievement,* 70 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1925.

[44] *Educational Leadership,* 1: 305–310, April, 1945.

[45] Ben D. Wood, "Basic Considerations," *Review of Educational Research,* 3: 7, February, 1933.

[46] A master's thesis summarized in Charles C. Peters' (Editor) *Abstracts of Studies in Education at The Pennsylvania State College, Part VI,* 1936, pages 27–28.

[47] Arthur E. Traxler, "The Use of Test Results in Secondary Schools," *Educational Records Bulletin,* 25: 8–9, 1938.

INFORMATION CONCERNING HOME

Name	Nat'l	Sep.	Dead	Religion	Language Spoken	Education
Father						
Mother						
Step-Father						
Step-Mother						

Siblings	Older	Younger	Date	Father's Occupation	Mother's Occupation
Brothers					
Sisters					
Others in Home					

Date	Home Address	Lives With	Phone

SUMMARY OF TEST DATA

Educational Test Data—Elementary						Educational Test Data—Junior High School						Educational Test Data—Senior High School					
Date	Name of Test	Form	Score	% or G. P.	Norm	Date	Name of Test	Form	Score	% or G. P.	Norm	Date	Name of Test	Form	Score	% or G. P.	Norm
*																	

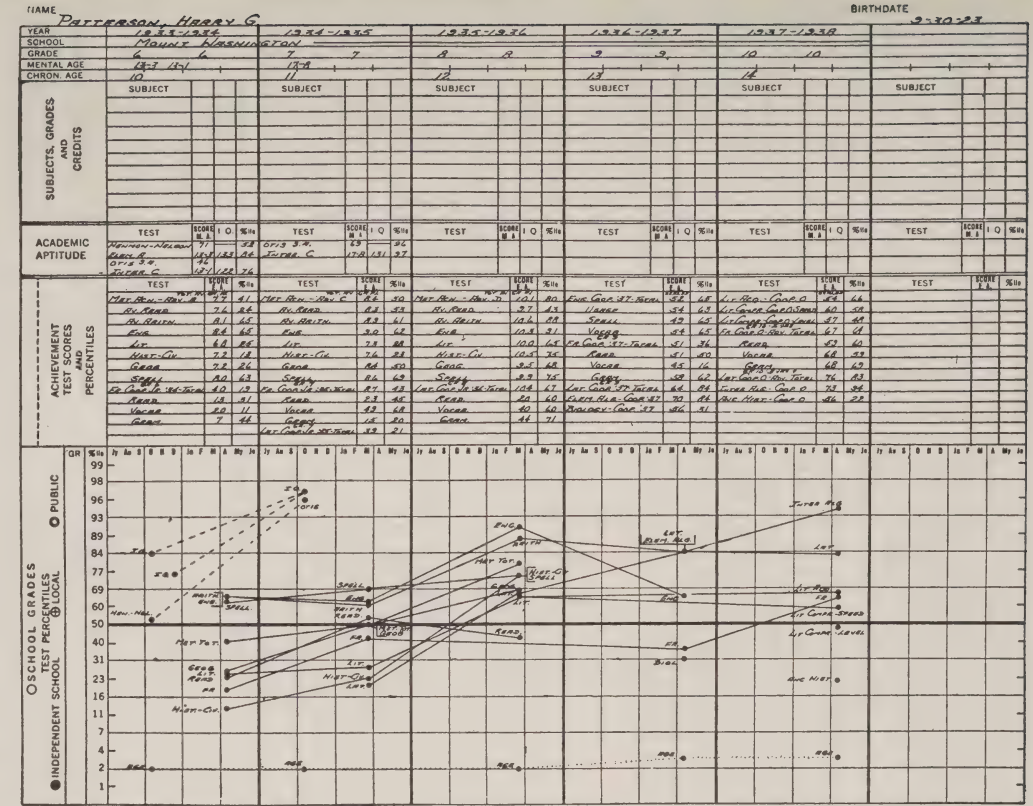
Tests of Mental Maturity							Personal and Social Adjustment Inventories							
Date	Name of Test	Grade	Form	CA	MA	IQ	Date	Grade	Name of Inventory	Results	Date	Grade	Name of Inventory	Results
*														

Significant Physical and Mental Health Information

Date		Date		Date	
*					

*More space may be allowed here, if needed.

Figure 7. Test Data Summary from the Cumulative Guidance Record of the Department of Supervision and Curriculum Development of the National Education Association.


SCHOOL GRADES ○ PUBLIC ⊕ LOCAL ● INDEPENDENT SCHOOL
TEST PERCENTILES
GR
%ile
99
98
96
93
89
84
77
69
60
50
40
31
23
16
11
7
4
2
1
Jy Au S O N D Ja F M A My Je
I.Q.
I.Q.
Otis
Hen.-Nel.
Arith.
Eng.
Spell.
Met. Tot.
Geog.
Lit.
Read.
Fr.
Hist.-Civ.
Spell.
Eng.
Arith.
Read.
Fr.
Lit.
Hist.-Civ.
Lat.
Hist.-Civ.
Geog.
Eng.
Arith.
Met. Tot.
Hist.-Civ.
Spell.
Geog.
Lat.
Lit.
Read.
Lit.
Elem. Alg.
Eng.
Fr.
Biol.
Inter. Alg.
Lat.
Lit. Acq.
Fr.
Lit. Compr. Speed
Lit. Compr. Level
Anc. Hist.
Age
Age
Age
Age
Age

NAME Patterson, Harry G.
BIRTHDATE 9-30-23

YEAR	1933-1934	1934-1935	1935-1936	1936-1937	1937-1938	
SCHOOL	Mount Washington					
GRADE	6 6	7 7	8 8	9 9	10 10	
MENTAL AGE	13-3 13-1	14-8				
CHRON. AGE	10	11	12	13	14	

SUBJECTS, GRADES AND CREDITS

SUBJECT				SUBJECT				SUBJECT				SUBJECT				SUBJECT				SUBJECT			

ACADEMIC APTITUDE

TEST	SCORE M. A.	I. Q.	%ile	TEST	SCORE M. A.	I. Q.	%ile	TEST	SCORE M. A.	I. Q.	%ile	TEST	SCORE M. A.	I. Q.	%ile	TEST	SCORE M. A.	I. Q.	%ile	TEST	SCORE M. A.	I. Q.	%ile
Henmon-Nelson	71		58	Otis S.A.	65		96																
Kuhl. A.	13-3	133	84	Inter. C	14-8	131	97																
Otis S.A.	46																						
Inter. C	13-1	122	76																				

ACHIEVEMENT TEST SCORES AND PERCENTILES

TEST	SCORE E. A.	%ile	TEST	SCORE E. A.	%ile	TEST	SCORE E. A.	%ile	TEST	SCORE E. A.	%ile	TEST	SCORE E. A.	%ile	TEST	SCORE E. A.	%ile
Met. Ach.-Bat. B	7.7	41	Met. Ach.-Bat. C	8.4	50	Met. Ach.-Bat. D	10.1	80	Eng. Coop. '37-Total	52	68	Lit. Acq.-Coop. O	54	66			
Av. Read	7.6	24	Av. Read	8.3	53	Av. Read	9.7	43	Usage	54	63	Lit. Compr. Coop. O-Speed	60	58			
Av. Arith.	8.1	65	Av. Arith.	8.9	61	Av. Arith.	10.6	88	Spell	49	65	Lit. Compr. Coop. O-Level	57	48			
Eng.	8.4	65	Eng.	9.0	62	Eng.	10.5	91	Vocab.	54	65	Fr. Coop. O-Rev. Total	67	61			
Lit.	6.8	26	Lit.	7.3	28	Lit.	10.0	65	Fr. Coop. '37-Total	51	36	Read	59	60			
Hist.-Civ.	7.2	13	Hist.-Civ.	7.6	23	Hist.-Civ.	10.5	75	Read	51	50	Vocab.	68	59			
Geog.	7.2	26	Geog.	8.6	50	Geog.	9.5	68	Vocab.	45	16	Gram.	68	69			
Spell	8.0	63	Spell	8.6	69	Spell	9.9	75	Gram.	58	62	Lat. Coop. O-Rev. Total	76	83			
Fr. Coop. Jr. '34-Total	40	19	Fr. Coop. Jr. '35-Total	87	43	Lat. Coop. Jr. '36-Total	104	67	Lat. Coop. '37-Total	64	84	Inter. Alg.-Coop. O	73	94			
Read	13	31	Read	23	45	Read	20	60	Elem. Alg.-Coop. '37	70	84	Anc. Hist.-Coop. O	56	22			
Vocab.	20	11	Vocab.	49	68	Vocab.	40	60	Biology-Coop. '37	56	51						
Gram.	7	44	Gram.	15	20	Gram.	44	71									
			Lat. Coop. Jr. '35-Total	39	21												

CUMULATIVE RECORD FOR INDEPENDENT SCHOOLS — EDUCATIONAL RECORDS BUREAU — 437 WEST 59TH STREET, NEW YORK, N. Y

Figure 8. A Cumulative Record in Graphical Form.

practical use. Schools with limited resources and little clerical assistance should be content with less elaborate record systems than those which may be feasible for larger and wealthier schools.

Reports to parents or public. Only a few schools make a systematic effort to keep the public informed regarding the educational progress of its schools. Traxler [48] found that only about one school in seven made a regular practice of using test results in reporting to parents, although only one in ten failed to do so altogether. Results of the testing program might very well be summarized before the Parent-Teacher Association, women's clubs, luncheon clubs, and similar organizations. Slides and charts, illustrating the nature of the tests, with analysis and interpretation of the records of typical pupils, would be instructive. The cumulative record cards are naturally of great value in conferences with parents regarding the educational program of their children. Hilkert [49] points out clearly how this may be done. A further discussion of the use of measurement in programs of public relations is given in Chapter XVIII.

SELECTED REFERENCES FOR FURTHER READING

Boynton, Paul L., *Intelligence: Its Manifestation and Measurement.* New York: D. Appleton-Century Company, 1933. Chapters VI, VII, and IX.

Buros, Oscar Krisen, *The Nineteen Thirty Eight Mental Measurements Yearbook.* New Brunswick, New Jersey: Rutgers University Press, 1938. 415 pages.

———, Succeeding *Mental Measurements Yearbooks* published by the author, Highland Park, New Jersey.

Greene, Harry A., Jorgensen, Albert N., and Gerberich, J. Raymond, *Measurement and Evaluation in the Elementary School.* New York: Longmans, Green & Company, 1942. Chapters VI and XXV.

Kandel, I. L., *Examinations and Their Substitutes in the United States.* New York: The Carnegie Foundation for the Advancement of Teaching, 1936. Chapter III.

Kelley, Truman Lee, *Interpretation of Educational Measurements.* Yonkers: World Book Company, 1927. Chapters II, III, and IV.

Lee, J. Murray, *A Guide to Measurement in Secondary Schools.* New York: D. Appleton-Century Company, 1936. Chapters II and XII.

McCall, William A., *Measurement.* New York: The Macmillan Company, 1939. Book Two.

National Committee on Cumulative Records, *Handbook of Cumulative Records.* Washington: U. S. Office of Education, 1944. 104 pages.

Nelson, M. J., *Tests and Measurements in Elementary Education.* New York: The Cordon Company, Inc., 1939. Chapters XII and XIII.

Orleans, Jacob S., *Measurement in Education.* New York: Thomas Nelson and Sons, 1937. Chapters VI, VII, and VIII.

---

[48] *Ibid.*, page 19.

[49] Robert N. Hilkert, "Parents and Cumulative Records," *Educational Record,* 21: 172–183, Supplement No. 13, January, 1940.

Smith, Eugene R., Tyler, Ralph W., and Evaluation Staff, *Appraising and Recording Student Progress.* New York: Harper & Brothers, 1942. Chapters VIII and IX.

Tiegs, Ernest W., *Tests and Measurements in the Improvement of Learning.* Boston: Houghton Mifflin Company, 1939. Chapter XVI.

Traxler, Arthur E., *Techniques of Guidance.* New York: Harper & Brothers, 1945. Chapters VIII–XII.

Woody, Clifford, and Sangren, Paul V., *Administration of the Testing Program.* Yonkers: World Book Company, 1932. 397 pages.

CHAPTER VIII

The Statistical Analysis of Test Results

A. The Importance of Educational Statistics

The purposes of the chapter. This chapter makes no pretense of being a complete treatment of statistical methods. It will attempt to present the minimum essentials only. The mathematics involved will not exceed that of sixth-grade arithmetic. The primary purposes of the chapter are two in number:

1. To develop in the student sufficient knowledge and skill to enable him to make the simple analyses of test scores that are essential to the intelligent interpretation and utilization of the results of measurement.
2. To give the student a vocabulary sufficient to enable him to read with understanding the extensive literature relating to measurement, much of it involving statistical terms, and to follow without difficulty or embarrassment the discussions at teachers' meetings and educational conferences of various kinds. A minimum of experience in computation is one of the best means of acquiring an adequate understanding of the terms and processes.

Mathews' study. A few years ago Mathews[1] made a comprehensive study of the content and purposes of the introductory course in educational measurement, and upon the basis of the consensus of experienced teachers, school administrators, and specialists in measurement, came to this conclusion:[2]

> Skill in the computation of only the simpler and more common statistical measures seems justified as an objective of the introductory course. These measures would include the central tendency, quartile points, percentiles, and the quartile deviation. The construction and interpretation of charts, tables, and distributions seem essential. Some degree of interpretative skill in dealing with other measures such as the standard deviation, correlation coefficient and reliability measures is desirable.

Dickey's study. Starting with the assumption that one of the most important objectives of the first course in tests and measurements is to develop the ability of the students to read the professional journals, Dickey[3] made a study of the statistical knowledge required to read the seven journals judged by the heads of departments in the New Jersey Normal School at Newark to be of greatest value to teachers in elementary schools. Dickey tabulated the statistical terms found in five issues of each of these journals for the years 1925–1926, 1928–1929, 1931–1932, and 1934–1935. Of the 898 articles included, 38.1 per cent used statistical measures varying in amount from 0 for *Progressive Education* to 81.2 for *The Journal of Educational Psychology.* The summary of terms occurring ten or more times is given in Table 13.

---

[1] C. O. Mathews, "The Introductory Course in Educational Measurements," *Educational Administration and Supervision,* 21: 431–447, September, 1935.

[2] *Ibid.,* page 447.

Table 13

SUMMARY OF TERMS OCCURRING TEN OR MORE TIMES IN 898 ARTICLES IN SEVEN PROFESSIONAL JOURNALS (AFTER DICKEY)

Measure	Frequency
Central Tendency:	
Mean	153
Median	101
Average	63
Variability:	
Standard Deviation	99
Range	52
Probable Error	44
Coefficient of Variation	11
Quartile Deviation	10
Correlation:	
Pearson *r*	148
Reliability:	
Difference	66
Standard Error of Difference	22
Reliability Coefficient	16
Standard Error of Mean	14
Difference Divided by Standard Error of Difference	14
Reliability	14
Critical Ratio	13
Probable Error	13
Probable Error of Difference	10
Miscellaneous:	
Quartiles	50
Frequency	32
Rank	16

---

[3] John W. Dickey, "Statistical Ability Necessary to Read Educational Journals," *Journal of Educational Psychology,* 27: 149–154, February, 1936.

Dickey concluded his study by suggesting the following additions to the list:[4]

> The *additions* are: The use of measure of central tendency as a general term to include the mean, the median, etc.; Spearman's Rho, because in many cases in education and psychology it is a question whether we have advanced any further than the ranking stage and therefore may have overworked the more refined correlation techniques; the coefficient of alienation (k), to be used to give meaning to the Pearson r; IQ; MA; CA; percentiles; cumulative frequency; norms (age and grade); normalcy (G); chances in one hundred that a difference is significant; scattergram; and a difference divided by the probable error of the difference.

The beginning student of measurement has doubtless found the discussion of these three studies difficult to follow because of the many unfamiliar terms encountered. This chapter will attempt to develop the basic concepts necessary for the intelligent understanding and interpretation of the professional literature of education, as well as the knowledge and skill required for the analysis of the results of a testing program.

Kittle[5] has reported an analysis of the statistical tools used in the educational research studies appearing in *The Journal of Educational Research* for the years 1920–1940, inclusive. She concluded that on the whole frequencies had remained "fairly constant," and were in line with those reported by Dickey. Kittle also noted a lack of consistency in the terminology employed and suggested that a greater degree of standardization of terms and symbols would make for ease of reading and understanding of research reports.

B. Classification and Tabulation

Before test scores or other quantitative data can be comprehended and interpreted, it is necessary to make an analysis of them. Table 14 gives a class record for a reading readiness test administered at the beginning of the school year. The scores appear in alphabetical order as they are recorded in the teacher's class roll book. However, the scores do not mean very much in this form. It is with some difficulty that we can tell whether Richard A, with a score of 90, for example, is a very superior or just an average pupil.

---

[4] *Ibid.*, page 154.

[5] Marian A. Kittle, "Trends in the Use of Statistical Tools in Educational Research Articles," *Journal of Educational Research,* 38: 34–46, September, 1944.

Table 14

A CLASS RECORD FOR A READING READINESS TEST

Pupil	Score
Richard A	90
Robert B	66
Barbara B	106
Charles B	84
Mildred C	105
Robert C	83
Robbin C	104
Diney D	82
Jim D	97
John D	97
Robert D	59
Don F	95
Larry F	78
Richard G	70
Warren H	47
Sylva H	95
Robert H	100
Grover H	69
Jack K	44
Clarence K	80
Jerome L	75
Mary M	75
Billy N	51
Nancy O	109
Carrie P	89
Ralph R	58
Billy S	59
William S	72
Gretta S	74
George S	75
Robert S	81
Jack S	71
Richard S	68
Mary S	112
Jean T	62
Richard W	91
Dolores W	93
Carl W	84

Table 15 gives the class record of an eighth-grade class on a general achievement test. In its present form it serves to measure very little. Some kind of analysis of the scores must be made before it is possible to secure a meaningful interpretation of the situation.

Rank order. Ordinarily the first step is to arrange the scores in order of size, usually from high to low. This is called an *ungrouped series.* In a small class this is sometimes all that is necessary.

Table 15

CLASS RECORD OF A GENERAL ACHIEVEMENT TEST FOR AN EIGHTH-GRADE CLASS

Pupil	CA	MA	EA	Read.	Arith.	Eng.	Lit.	Hist.	Geog.	Spell.
1	13–5	17–4	15–8	13–4	15–6	16–3	16–3	16–3	17–0	15–7
2	12–3	18–2	15–6	15–3	15–5	16–1	16–10	15–9	17–0	15–4
3	12–11	16–9	15–5	15–5	14–10	16–3	16–10	16–6	15–3	14–0
4	13–4	15–5	15–3	15–4	13–11	15–7	16–3	16–0	15–3	13–4
5	11–8	13–9	15–3	15–1	15–0	15–3	15–11	15–3	16–5	14–3
6	13–4	15–11	15–2	14–11	14–8	15–5	15–11	15–11	17–0	12–6
7	12–10	15–5	15–1	14–5	15–1	15–1	17–0	15–1	17–0	15–1
8	13–8	16–1	15–0	14–11	15–2	14–8	16–10	15–0	16–7	13–11
9	13–9	15–1	14–11	15–1	13–6	14–8	15–11	14–11	15–9	13–2
10	13–1	13–9	14–8	14–6	14–10	14–8	15–6	15–2	16–5	13–11
11	13–11	15–7	14–8	14–7	13–6	15–1	17–0	16–7	17–0	13–0
12	13–1	14–8	14–8	14–3	14–7	14–8	15–8	15–0	17–0	13–2
13	13–5	15–1	14–8	14–8	14–8	14–8	15–6	14–5	15–11	13–6
14	13–1	13–10	14–7	14–8	14–11	14–9	14–7	13–10	15–2	14–4
15	13–1	16–5	14–7	15–0	13–4	15–1	15–10	13–10	15–2	13–0
16	13–5	12–10	14–5	13–8	14–1	13–9	15–8	15–4	17–0	12–0
17	14–11	15–0	14–3	14–0	13–2	15–6	15–4	15–6	15–2	13–6
18	12–8	13–9	13–11	13–10	13–10	14–3	14–3	13–8	14–3	12–8
19	13–4	14–9	13–11	13–8	13–0	14–7	16–2	16–7	17–0	12–4
20	12–11	13–8	13–9	12–3	14–7	13–11	13–0	14–0	13–9	12–9
21	13–6	12–10	13–8	12–10	13–8	14–9	15–5	13–8	14–9	12–7
22	13–0	13–1	13–8	12–11	13–8	14–4	14–7	14–4	14–7	11–8
23	13–5	14–7	13–8	12–9	13–5	14–5	15–5	14–6	15–0	12–6
24	13–1	15–7	13–6	13–8	13–6	10–8	15–0	12–11	12–10	14–8
25	12–8	13–10	13–5	13–6	13–4	14–6	13–2	13–4	13–4	14–7
26	13–0	12–7	12–11	12–9	12–9	14–5	14–3	12–11	13–11	12–9
27	13–5	12–2	12–10	12–0	13–2	13–7	12–10	13–4	13–10	12–1

Table 16 gives the same scores as Table 14 arranged in order of size. This table also shows the *rank order* of the pupils and the scores tabulated without grouping. It is now easy to see that Richard A's score of 90 gives him a rank of thirteen in a class of thirty-eight, or about one third of the way from the top. In a similar manner, it is easy to interpret each of the other scores in terms of rank. But this method, especially in classes of twenty or more pupils, is likely to prove unsatisfactory. Note, for example, that two pupils make a score of 97. Since it is not correct to say that one ranks higher than the other, it is necessary to assign them fractional ranks. As there are six pupils who rank higher, the next two ranks, 7 and 8, are averaged, which gives 7.5. In like manner the average of ranks 9 and 10 is 9.5, and so on for the other pupils who make the same scores. Since there are three pupils each of whom makes a score of 75, and there are twenty-one pupils who rank above this score, the average of the next three ranks, 22, 23, and 24, is 23, which is the rank assigned each of the scores of 75. In addition to the fact that

Table 16

READING READINESS SCORES ARRANGED IN ORDER OF SIZE, RANK ORDER, AND TABULATED WITHOUT GROUPING

Order of Size	Rank Order	Tabulated Without Grouping	
		Score	*Frequency*
112	1	112	1
109	2	109	1
106	3	106	1
105	4	105	1
104	5	104	1
100	6	100	1
97	7.5	97	2
97	7.5	95	2
95	9.5	93	1
95	9.5	91	1
93	11	90	1
91	12	89	1
90	13	84	2
89	14	83	1
84	15.5	82	1
84	15.5	81	1
83	17	80	1
82	18	78	1
81	19	75	3
80	20	74	1
78	21	72	1
75	23	71	1
75	23	70	1
75	23	69	1
74	25	68	1
72	26	66	1
71	27	62	1
70	28	59	2
69	29	58	1
68	30	51	1
66	31	47	1
62	32	44	1
59	33.5		
59	33.5	Total	38
58	35		
51	36		
47	37		
44	38		

time and trouble are required to determine these ranks, the list is long and unwieldy to handle, and is completely inadequate for making comparisons with other classes which may be much larger or much smaller.

The frequency table or distribution. The way out of the difficulty is to arrange the scores into a table. Such a process is called *tabulation.* The table itself is called a *frequency table,* a *frequency distribution,* or merely *distribution.* The third and fourth columns of Table 16 show the simplest form of a distribution. Such a distribution consists of two columns; the various scores are arranged in one column in order of size, and opposite each score is recorded in the other column the number of times it occurs. Each entry in the second column is called a *frequency,* abbreviated f, and the total is represented by N.

It is usually necessary, however, to carry the process one step further. As a rule, there is such a wide range of scores that it is desirable to classify them into groups according to size. Each group is called a class. This arrangement is usually referred to as the *grouped frequency distribution.* While there is no absolutely fixed rule for the number of classes, it is usually advisable to make *not fewer than ten classes nor more than twenty.* To make fewer than ten classes is to run the risk of seriously affecting the accuracy of the results, while to make more than twenty classes is to produce a table that is inconvenient to handle. It has been found that a table of between ten and twenty classes is convenient to handle and does not seriously distort the accuracy of the results.

Making the frequency table. There are four steps in making the ordinary frequency table or grouped frequency distribution. These are illustrated in Table 17, using the scores given in Table 16.

1. *Determine the range,* which is the difference between the highest score and the lowest. Of these scores, the highest is 112 and the lowest is 44, which gives a range of 68.

2. *Select the class interval,* which is the size of the groups into which the scores are to be classified. To do this, divide the range by 10, which gives the largest group, or class interval, that can be used; and by 20, which gives the smallest class interval that can be used. In this case, $68 \div 10 = 6.8$ and $68 \div 20 = 3.4$. Since it is impractical to use any class interval except a whole number, the fractions are disregarded and the next highest whole number is taken. The class interval must, therefore, be somewhere between 4 and 7. A class interval of 4, 5, 6, or 7 might be used. Of the available class intervals it is best to choose the one which is most convenient to use; in this case 5 is probably the best choice.

3. *Determine the limits of the classes.* The table must, of course, be long enough to include the highest score and the lowest score. To facilitate tabulation start each class with a multiple of the class interval. If the highest class starts with 110, which is a multiple of 5, it will accommodate the highest score, 112. Each succeeding class will drop back 5 points below the one just above it. The

next class will start at 105, the next at 100, and so on, till the lowest one, which starts at 40, is reached.

4. *Make the tabulation.* A short vertical line is drawn for each score opposite the class where it falls. To make a tabulation it is not necessary to have all the scores arranged in order, and it is not advisable to do so, for this process usually requires more time than the tabulation itself. In the original alphabetical list the first score is 90. In the "tabulation" column opposite the class which begins with 90, a vertical line is drawn to indicate the score. The next score is 66. This

TABLE 17

AN ILLUSTRATION OF THE PROCESS OF MAKING A FREQUENCY TABLE

<table>
<tr><th>Original Scores</th><th>Steps in Making the Table</th></tr>
<tr><td rowspan="3">90
66
106
84
105
83
104
82
97
97
59
95
78
70
47
95
100
69
44
80
75
75
51
109
89
58
59
72
74
75
81
71
68
112
62
91
93
84</td>
<td>Step 1. Determining the range.<table>
<tr><td>Highest Score</td><td>112</td></tr>
<tr><td>Lowest Score</td><td>44</td></tr>
<tr><td>Difference (range)</td><td>68</td></tr>
</table></td></tr>
<tr><td>Step 2. Selecting the class interval.
68 ÷ 10 = 6.8, largest class interval possible.
68 ÷ 20 = 3.4, smallest class interval possible.
(5 chosen because of convenience in tabulation).</td></tr>
<tr><td>Steps 3 and 4. Determining the limits of the classes and making the tabulation.<table>
<tr><th><i>Limits of Classes</i></th><th><i>Tabulation</i></th><th><i>Frequency (f)</i></th></tr>
<tr><td>110–114</td><td>/</td><td>1</td></tr>
<tr><td>105–109</td><td>///</td><td>3</td></tr>
<tr><td>100–104</td><td>//</td><td>2</td></tr>
<tr><td>95–99</td><td>////</td><td>4</td></tr>
<tr><td>90–94</td><td>///</td><td>3</td></tr>
<tr><td>85–89</td><td>/</td><td>1</td></tr>
<tr><td>80–84</td><td>~~////~~ /</td><td>6</td></tr>
<tr><td>75–79</td><td>////</td><td>4</td></tr>
<tr><td>70–74</td><td>////</td><td>4</td></tr>
<tr><td>65–69</td><td>///</td><td>3</td></tr>
<tr><td>60–64</td><td>/</td><td>1</td></tr>
<tr><td>55–59</td><td>///</td><td>3</td></tr>
<tr><td>50–54</td><td>/</td><td>1</td></tr>
<tr><td>45–49</td><td>/</td><td>1</td></tr>
<tr><td>40–44</td><td>/</td><td>1</td></tr>
</table></td></tr>
</table>

falls in the class which begins at 65, so a line is made there. In the same way a line is placed in the column opposite the appropriate class. For the fifth score in each class a diagonal line is drawn across the other four. This makes it easier to count the frequency in each class. The frequency column, abbreviated *f*, gives the number of scores that fall in each class.

The finished table omits the steps by which it was made. In the simplest form of the table only two columns occur, the first of which shows the various classes, usually arranged in descending order, and the second of which shows the frequency or the number of scores in each class. When two or more schools or grades are to be compared, it is usually best to include all the data in the same table. In that case there will be a column for the classes into which the scores are grouped and one for each of the schools or grades being compared. Table 18 shows a frequency table which combines the record of six schools on a certain test.

Table 18

DISTRIBUTION OF READING READINESS SCORES FOR SIX SCHOOLS IN A CERTAIN CITY

Score	School A	School B	School C	School D	School E	School F
120–124				1		
115–119						
110–114			1			
105–109			3		2	2
100–104		3	2	2	5	3
95–99		6	4	4	4	5
90–94	5	2	3	5	6	10
85–89	4	4	1	4	4	1
80–84	2	3	6	6	4	8
75–79	10	5	4	4	1	2
70–74	6	2	4	7	6	4
65–69	9	4	3		4	1
60–64	4	5	1	2	1	
55–59	1		3		1	
50–54	1		1			
45–49	1		1			
40–44			1	2	2	
35–39	1	1				
30–34		2				
25–29		1				
20–24						
15–19						
10–14	1					
N	45	38	38	37	40	36

The form of the table. A few words may be said about the mechanical make-up of the table as it occurs in printed or typed form. Each table bears a number. Either Roman or Arabic numerals may be employed, but the latter seem to be increasingly favored. The table number may be centered above the title of the table, or it may be given at the beginning of the title. The title itself appears in capitals, and with no period or other punctuation mark after it. The table usually starts with two horizontal lines and ends with a single horizontal line. Another horizontal line separates the column headings from the body of the table, and other horizontal lines separate any summarizing measures which may be given under the table proper. Vertical lines may be used to separate the columns, but usually no lines are drawn along the margins of the page. It is considered good form to avoid abbreviations in the table whenever possible, and to make the title and headings full enough to indicate clearly the contents of the table.

Table 19

THE CHRONOLOGICAL, EDUCATIONAL, AND MENTAL AGES FOR AN EIGHTH-GRADE CLASS

Pupil	Ages Expressed in Months		
	Chronological (CA)	*Educational (EA)*	*Mental (MA)*
A	150	188	208
B	147	186	218
C	155	185	201
D	160	183	185
E	141	183	165
F	160	182	191
G	154	181	185
H	164	180	193
I	165	179	181
J	157	176	165
K	167	176	187
L	157	176	176
M	161	176	180
N	157	175	166
O	158	174	197
P	161	173	154
Q	179	171	180
R	152	167	165
S	160	167	177
T	156	165	164

A two-way table or scattergram. It is sometimes useful to compare pupils' scores on two measures at the same time. To do this a two-way table, or scattergram, is made. Table 19 shows the chronological, educational, and mental ages for a certain eighth-grade class. It would, of course, be possible to make three separate tables of these data. But a single two-way distribution, shown in Table 20, makes possible a comparison between the educational age and the mental age of these pupils. The educational age, grouped into class intervals of two months, is shown in the horizontal rows; the mental age, grouped into class intervals of six months, is shown in the vertical columns. For example, Pupil A, with an EA of 188 months and a MA of 208 months, falls in the top row, or 188- class, and in the third column from the right, or 204- class. In like manner, the horizontal position of each pupil in the distribution shows his EA, and the vertical position shows his MA. A tendency will be observed for the scores to arrange themselves in a diagonal pattern from upper right to the lower left. This means that, in general, pupils who are high in EA are high in MA, and pupils who

TABLE 20

A TWO-WAY DISTRIBUTION OF EDUCATIONAL AGE AND MENTAL AGE FOR AN EIGHTH-GRADE CLASS

	Mental Age in Months												EA
Educational Age in Months	150-	156-	162-	168-	174-	180-	186-	192-	198-	204-	210-	216-	Freq.
188-										A			1
186-												B	1
184-									C				1
182-			E			D	F						3
180-						G		H					2
178-						I							1
176-			J		L	M	K						4
174-			N					O					2
172-	P												1
170-						Q							1
168-													0
166-			R		S								2
164-			T										1
MA Frequency	1	0	5	0	2	5	2	2	1	1	0	1	20

are low in EA are low in MA. A few exceptions stand out, however. For example, Pupil P, who is lowest in MA, is in the fifth row from the bottom in EA. When the identification of individual pupils is unimportant, the totals only are entered in the appropriate squares.

C. Measures of Average or Central Tendency

The concept of average. In order to comprehend fully or to interpret the data in the table, it is necessary to make an analysis of it. Characteristic of most frequency tables is a tendency for the scores to bunch or concentrate somewhere near the center. The first and most important measure in statistical analysis is, therefore, the location of the point on the scale where the scores tend to group themselves. This measure is known as the *average*, or *central tendency.* It is that value which typifies, or best represents, the whole distribution.

It will be recalled that the norm on a standard test is the score made by the typical pupil of a given age or grade. In other words, the norm is the average score made by a large and representative group of pupils. When standard tests are given, the teachers are usually interested in knowing how the grade or school compares with the norms on the test. All comparisons with norms are in terms of averages. If the average score of a certain grade, for example, is the same as the norm on the test, it may be regarded as a typical or average grade as measured by that particular test. But, if the average of this grade is above the norm on the test, it is a better than average grade; and, if the average of this grade is below the norm on the test, it is a poorer than average grade.

Often, also, it is important to compare one grade or school with another. For example, one might wish to know which of several schools made the best record on a certain test, and which the poorest. To determine this, all that is necessary is to compute an average for each grade or school, and then to note which one has the highest average and which one has the lowest average. In other words, that school is best which *on the average* makes the highest score, and that school is poorest which *on the average* makes the poorest score.

Statisticians employ three common averages. These are the mode, or inspectional average; the median, or counting average; and the mean, or computed average. The meaning of each of these will now be considered.

The mode. The commonest score in a group is called the *mode.* It is obtained by inspection. In Table 16 on page 221 the mode is 75, because more pupils in this grade make that score than any other. The mode is not a very trustworthy average, however, especially with small groups. In this case the changing of a single score would shift the mode decidedly. If one of the pupils who made 75 had made 59, the mode would drop to 59, since more pupils would then have made that score than any other. In like manner, if one of the pupils who made 75 had made 97, the mode would rise to 97. Largely because of its fickleness, the mode is not highly regarded as a measure of average or central tendency.

The median. Perhaps the most widely used average in educational measurement is the *median.* The median is the mid-point in the distribution, or that point which divides the distribution into halves. Sometimes in an ungrouped series the *mid-score* is used instead of the median. Strictly speaking, when N is an even number, there is no mid-score. In that case, it is customary to average the middle pair of scores. For example, in Table 16 there are 38 pupils, 19 of whom make scores of 80 or less and 19 of whom make scores of 81 or more. The mid-score is then assumed to be the average of 80 and 81, the middle pair of scores, or 80.5. The terms *median* and *mid-score* are often used interchangeably, but there is a clear distinction between them: the median is a *point,* and the mid-score is a *score.* The latter should be used only for small classes, where the scores are arranged in order of size rather than in table form.

Table 21 illustrates the process of locating the median in a frequency table. The median is often described as the counting average, and it will be noted that counting does occupy a prominent place in its location. The steps may be summarized as follows:

1. *Obtain ½N.* That is, divide the total of the frequencies by 2. Here ½N or $\frac{N}{2} = \frac{38}{2} = 19$.

2. *Locate the approximate median.* Beginning at the low end of the distribution, count up the frequency column as far as possible without passing $\frac{N}{2}$, obtained in step 1. In this case the frequencies $1 + 1 + 1 + 3 + 1 + 3 + 4 + 4$ give a total of 18. This is as far as we can go, for to include the next frequency, 6, would carry us too far, or beyond $\frac{N}{2}$, which is 19. The approximate median then is 80, the lower limit of the class containing the median.[6]

3. *Determine the correction needed.* From $\frac{N}{2}$ subtract the total obtained in step 2. In this case, $19 - 18 = 1$. This shows that one more score or unit is needed to obtain the required half. And this score must come out of the next

---

[6] Strictly speaking, with certain types of data the lower limit would really be 79.5, rather than 80. Intelligence quotients would be an example. Since quotients are recorded to the nearest whole number, the minimum value of 80 would be 79.5. The best procedure, however, is to take the scores at their face value in all computations.

class, the 80–84 class, where there is a frequency of 6. That is, we must go ⅙ of the distance into the next class. As the class interval is 5, this means ⅙ of 5, or .83. The correction is then .83.

4. *Obtain the true median.* This is done by adding the correction to the approximate median. In this case 80 + .83 = 80.83, the median.

Table 21

THE PROCESS OF LOCATING THE MEDIAN

Frequency Table			Steps in the Process
	f		
110–114	1		Step 1. Obtaining $\frac{N}{2}$. $\frac{N}{2} = \frac{38}{2} = 19.$
105–109	3		
100–104	2		
95–99	4		
90–94	3		
85–89	1		Step 2. Locating approximate median. 1 + 1 + 1 + 3 + 1 + 3 + 4 + 4 = 18. This takes us up to 80, which is the *approximate median.*
80–84	6		
75–79	4	18	
70–74	4		
65–69	3		
60–64	1		
55–59	3		Step 3. Determining the correction. 19 — 18 = 1. $\frac{1}{6} \times 5 = .83$, the *correction.*
50–54	1		
45–49	1		
40–44	1		
N	38		
			Step 4. Locating the true median. 80 + .83 = 80.83, the true median. That is, the true median is the approximate median plus the correction.

It is, of course, possible to count down instead of up the frequency column and obtain the same result. But the correction would then be ⅚ of 5, or 4.17; and it would be taken away from 85, the upper limit of the 80–84 class, which includes all scores between 80 and 85. While counting down gives the same result as counting up, the writer has found that beginning students in measurement make fewer mistakes when counting up. The counting-up method is, therefore, recommended; the other is a useful check.

Students trained in algebra may be interested in the formula for the median. Such students may prefer to think of the above process as substituting in a formula, although it is not at all necessary to do so. The formula may be written:

$$Md = l + \frac{\frac{1}{2}N - S_b}{f}\, i.$$

In this formula

$l =$ lower limit of class containing median, and may be called the approximate median.
$\frac{1}{2}N =$ one half the number of cases, or one half of the sum of the f column.
$S_b =$ sum of frequencies below the class containing the point desired.
$f =$ frequency in class containing median.
$i =$ class interval used in the table.

The median is often used as a reference point for describing the location of individual pupils in a distribution. A pupil in the higher half is said to be "above the median," and one in the lower half is said to be "below the median." Other points in the distribution are used in a similar manner. For example, *quartiles* divide the distribution into fourths, and *deciles* divide it into tenths. A pupil in the highest fourth is said to be "above Q_3," and one in the lowest fourth is said to be "below Q_1." The position of a certain pupil may be still more accurately described by indicating the percentage of pupils who fall below him. The points that divide a distribution into 100 equal divisions, or per cents, are called *percentiles.* The general formula is the same for all such points in the distribution. If x is the proportion of the distribution that falls below the desired point, P, the general formula is:

$$P = l + \frac{xN - S_b}{f} i.$$

The only difference between the formulas for the quartiles, deciles, percentiles, and the like is in the proportion of the distribution that is cut off. For example,

$$Q_1 = l + \frac{\frac{1}{4}N - S_b}{f} i,$$

$$Q_3 = l + \frac{\frac{3}{4}N - S_b}{f} i,$$

$$P_{10} = l + \frac{10/100N - S_b}{f} i,$$

and so on. Table 23 on page 235 illustrates the computation of the quartiles.

The mean. The most familiar average is the *mean,* often called the *arithmetic mean.* In fact, this measure is in such common use that the ordinary person regards it as *the* average, because it is the only average he knows anything about. When the term "average" is met with in ordinary conversation or the newspaper in such statements as "average temperature," "average rainfall," "average yield of corn and wheat," "average price," and the like, it is almost certain to be the mean that is meant. The mean can be computed merely by obtaining the sum of the measures and dividing by their number. The measure so obtained is then the value that each individual would have if all shared equally.

When the scores are few in number or in an ungrouped series, the simplest process of computing the mean is the one described above; that is, the scores are first added and then this sum is divided by the number of scores. This is known as the "long" method. For example, the sum of the 38 scores in Table 17 is 3,050, and $3{,}050 \div 38 = 80.3$ Since Σ means "the sum of," the formula may be written:

$$M = \frac{\Sigma f}{N}.$$

When the scores are sufficiently numerous to justify the use of the frequency table, the so-called "short" method of computing the mean is more convenient.

$$M = M' + \frac{\Sigma fd}{N} i.$$

In this formula

M = true mean,
M' = assumed mean,
Σfd = sum of frequencies multiplied by their respective deviations,
N = total number of frequencies,
i = class interval.

Since $\frac{\Sigma fd}{N}\, i$ is the formula for the correction, abbreviated c, the formula may be shortened to $M = M' + c$.

The method of computing the mean by the short formula is illustrated in Table 22. The steps in the process are as follows:

Step 1. *Assume a mean.* This is taken at the mid-point of some class. Any class may be taken, but it is usually best to choose one near the center of the distribution. In this case the assumed mean is taken at the mid-point of the 80–84 class, whose lower limit is 80 and whose upper limit is 85. The midpoint is 82.5.

Step 2. *Lay off the deviations from the assumed mean.* The plus deviations indicate how many classes various frequencies are above the assumed mean, and minus deviations indicate how many classes various frequencies are below the assumed mean. This column is headed d.

Step 3. *Multiply each f by its corresponding d.* This column is headed fd. The first product is $1 \times 6 = 6$, the second is $3 \times 5 = 15$, and so on.

Step 4. *Obtain the algebraic sum of the fd column.* Note that the sum of the + values is 48 and the sum of the − values is −61. The algebraic sum is −13. Had the + values exceeded the − values, the sum would have been +. This is called Σfd.

Step 5. *Determine the correction.* To do this divide the sum of the *fd* column by *N*, and multiply by the class interval.
$-13 \div 38 = -.34$, the correction in terms of classes.
$-.34 \times 5 = -1.7$, the correction in terms of score units, which is the desired value.

Step 6. *Obtain the true mean by adding the correction to the assumed mean.* Here the assumed mean is 82.5 and the correction is −1.7. The minus sign in the correction indicates that the assumed mean is too high, and must be corrected downward. $82.5 + (-1.7) = 80.8$, the true mean. Note that the true mean is the algebraic sum of the assumed mean and the correction. The sign of the correction must be observed.

TABLE 22

THE PROCESS OF COMPUTING THE MEAN

COMPUTATION				STEPS IN THE PROCESS
	f	*d*	*fd*	
110–114	1	+6	+ 6	Step 1. Assuming a mean. 82.5 is taken as the assumed mean. Two parallel lines indicate the class in which it is located.
105–109	3	+5	+15	
100–104	2	+4	+ 8	
95–99	4	+3	+12	
90–94	3	+2	+ 6	Step 2. Laying off deviations from assumed mean. This is the column headed *d*.
85–89	1	+1	+ 1	
80–84	6	0		
75–79	4	−1	− 4	
70–74	4	−2	− 8	Step 3. Multiplying each *f* by its *d* This column is headed *fd*.
65–69	3	−3	− 9	
60–64	1	−4	− 4	
55–59	3	−5	−15	Step 4. Obtaining algebraic sum of the *fd* column. Sum of + values is 48. Sum of — values is —61. Algebraic sum is —13. This is Σfd.
50–54	1	−6	− 6	
45–49	1	−7	− 7	
40–44	1	−8	− 8	
	38		+48	
			−61	
		38)	−13	Step 5. Determining the correction. $-13 \div 38 = -.34$. $-.34 \times 5 = -1.7$, correction in score units, which is the form needed.
		$c =$	−.34	
		$i =$	5	
		$ci =$	−1.70	
		$M' =$	82.5	
		$M =$	80.8	
				Step 6. Adding correction to assumed mean. Assumed mean 82.5 Correction —1.7 True mean 80.8

It will be recalled that the mean when computed by the "long" method was 80.3, or .5 less than when computed by the "short" method. This variation is due to the fact that the former method is based upon the actual value of the scores, whereas the latter method is based on the assumption that the mid-point of each class is the average for all the scores in that class, an assumption which is usually only approximately true. As a rule, the difference in the result is so slight as to make no practical difference in the interpretation of the situation.

What average is best? As a rule, the mean is regarded as the best average, and the mode is certainly the poorest. The mean, however, is greatly influenced by extreme scores, and whenever it is desired to avoid this influence, the median is to be preferred. As such situations often arise in educational measurement, the median is widely used as a measure of average. For example, if the test is too difficult, there may be several zero scores; and if the test is too easy, there may be several perfect scores. But in neither case are the pupils at the extremes correctly measured. The median is in such situations the best average to use. The median is also much easier to find. In fact, with test scores the mean is rarely ever sufficiently more accurate than the median to justify the additional labor required to compute it.

D. Measures of Variability or Scatter

Meaning of variability. No distribution is completely described by its average or central tendency. Two classes in a school might have the same average intelligence and yet be very unlike. The members of one class might vary all the way from feeble-mindedness to the genius level, while all the members of the other group may rate as normal. Obviously, these two classes present different instructional problems, because they differ in *variability.* Variability is the extent to which the scores tend to scatter or spread above and below the average. It is clearly important to have some convenient method of indicating the variability of a group. The second problem in the statistical analysis of test data is to determine the variability of the scores. There are three common measures of variability: namely, the range, the quartile deviation, and the standard deviation. All these measures represent distances rather than points, and the larger they are the greater the variability or scatter of the scores.

The range. The *range* has already been referred to as the distance between the lowest and the highest scores. The range, however, is a very untrustworthy measure of variability. It can be seen that the shift in a single score may greatly alter the range, and so materially increase or reduce the apparent variability of the group. School A and School D in Table 18 illustrate this possibility.

The quartile deviation. A measure of variabiltiy that avoids being unduly influenced by a few extreme scores is the *quartile deviation,* or Q. This is one half the distance between the first and third quartiles. For this reason, it is often referred to as the semi-interquartile range. Since 25 per cent of the scores fall below the first quartile, or Q_1, and 25 per cent of the scores exceed the third quartile, or Q_3, the interquartile range is the range of the middle 50 per cent of the scores. The whole interquartile range might be used to express the variability of the group, but it is customary to take only half this distance. The formula used is

$$Q = \frac{Q_3 - Q_1}{2}.$$

Table 23 illustrates the computation of Q. It will be observed that the process of locating quartiles and other similar points is very similar to that of locating the median. In fact, the only difference at all is in the first step, where the fractional part of N always indicates the proportion of the distribution which falls below the desired point; that is, for Q_1 it is $\frac{1}{4}N$ and for Q_3 it is $\frac{3}{4}N$. The process is simple, and represents three steps as follows:

1. *Compute* Q_1. To begin with, ¼ of 38 is 9.5. The next three steps in locating this point are exactly the same as those in locating the median.
2. *Compute* Q_3. Here the first step is to take $\frac{3}{4}N$, and ¾ of 38 is 28.5. The other three steps are identical with those in locating the median and Q_1.
3. *Substitute in the formula.* Q_1 is 69.17, and Q_3 is 95.63. The difference between them is 26.46. Half of this difference is 13.23, the value of Q.

The interpretation of Q and other measures of variability is a relative matter. Whether a Q of 13.23 is to be considered great or small depends upon the magnitude of comparable measures for other groups using the same test.

The standard deviation. A third measure of variability, which has many uses in educational measurement, is the *standard deviation.* This measure is usually represented by the Greek letter σ, or sigma. For this reason the standard deviation is often referred to as "sigma." It is defined as the square root of the mean of the squares of the deviations of the scores from their mean. It may also be defined as that distance above and below the mean that in a normal distribution includes 68.26 per cent of the scores, or approximately two thirds.

TABLE 23

THE PROCESS OF COMPUTING THE QUARTILE DEVIATION, OR Q

FREQUENCY TABLE

	f	
110–114	1	
105–109	3	
100–104	2	
95–99	4	
90–94	3	28
85–89	1	
80–84	6	
75–79	4	
70–74	4	
65–69	3	
60–64	1	7
55–59	3	
50–54	1	
45–49	1	
40–44	1	
N	38	

STEPS IN THE PROCESS

Step 1. Computing Q_1.
$\frac{1}{4}N = \frac{1}{4}$ of $38 = 9.5$.
Counting up: $1 + 1 + 1 + 3 + 1 = 7$;
approximate Q_1 is 65.
$9.5 - 7 = 2.5$; $\frac{2.5}{3} \times 5 = 4.17$, correction.
$65 + 4.17 = 69.17$, Q_1.

Step 2. Computing Q_3.
Counting up: $1+1+1+3+1+3+4+4+6+1+3 = 28$;
$\frac{3}{4}N = \frac{3}{4}$ of $38 = 28.5$.
approximate Q_3 is 95.
$28.5 - 28 = .5$; $\frac{.5}{4} \times 5 = .63$, correction.
$95 + .63 = 95.63$, Q_3.

Step 3. Substituting in formula.
Formula: $Q = \frac{Q_3 - Q_1}{2}$.
Substituting: $Q = \frac{95.63 - 69.17}{2} = 13.23$.

The formula for the standard deviation when computed from an assumed mean for scores in a frequency table is

$$\sigma = {}_i\sqrt{\frac{\Sigma fd^2}{N} - c^2}.$$

This is the usual situation, and the process is illustrated in Table 24. It can be seen that the only new term is Σfd^2, which is the sum of the frequencies times the squares of their respective deviations. The steps needed for computing the standard deviation, in addition to those needed in computing the mean by the short method, are three. The entire process is as follows:

A. Steps as in computing the mean.

Step 1. *Assume a mean.* Here it is 82.5.

Step 2. *Lay off the deviations above and below the assumed mean.*

Step 3. *Multiply each f by its d.*

Step 4. *Obtain Σfd, the algebraic sum of fd column.* Here it is -13.

Step 5. *Determine the correction.* $-13 \div 38 = -.34$. Note that the correction here is not multiplied by the class interval.

Table 24

THE PROCESS OF COMPUTING THE STANDARD DEVIATION

Computation

	f	d	fd	fd^2
110–114	1	+6	+ 6	36
105–109	3	+5	+15	75
100–104	2	+4	+ 8	32
95–99	4	+3	+12	36
90–94	3	+2	+ 6	12
85–89	1	+1	+ 1	1
80–84	6			
75–79	4	−1	− 4	4
70–74	4	−2	− 8	16
65–69	3	−3	− 9	27
60–64	1	−4	− 4	16
55–59	3	−5	−15	75
50–54	1	−6	− 6	36
45–49	1	−7	− 7	49
40–44	1	−8	− 8	64
	38		+48	479
			−61	
		38)	−13	

$c = -.34.$

$c^2 = .1156.$

$\sigma = 5\sqrt{479/38 - (-.34)^2},$

$= 5\sqrt{12.6053 - .1156},$

$= 5\sqrt{12.4897},$

$= 5 \times 3.53 = 17.65.$

Steps in the Process

Steps as in computing mean:

Step 1. Assume a mean (82.5).

Step 2. Lay off deviations from assumed mean.

Step 3. Multiply each f by its d.

Step 4. Obtain Σfd, the algebraic sum of fd column. Here it is —13.

Step 5. Divide by N to determine the correction. Here, $-13 \div 38 = -.34$.

Additional steps:

Step 6. Prepare fd^2 column. Each entry is product of d and fd opposite it.

Step 7. Obtain Σfd^2. This is merely the sum of the fd^2 column. Here it is 479.

Step 8. Substitute in formula.

Formula:

$$\sigma = i\sqrt{\frac{\Sigma fd^2}{N} - c^2}.$$

B. Additional steps required.

Step 6. *Prepare the fd^2 column.* Each entry in this column is the product of a d and the fd opposite it.

Step 7. *Obtain Σfd^2, the sum of fd^2 column.* All values in fd^2 column are positive, since negative deviations are squared. The sum is 479.

Step 8. *Substitute in the formula for σ.* Note that c^2 is always subtracted from Σfd^2, even when c is negative. The final result is in terms of score units, not classes. A common error is the failure to multiply by the class interval.

Practical uses of the standard deviation. The standard deviation is the most important measure of the variability of test scores. A small standard deviation means that the group has small variability, or is relatively homogeneous, while a large standard deviation means the opposite condition. It also has certain other important uses.

The position of a pupil in a distribution is often represented in terms of standard deviation units. In the distribution used in

Table 24, where the mean is 80.8 and the standard deviation is 17.65, a pupil whose score is 98 is said to be one standard deviation above the mean, and the score is written $+ 1\sigma$. In like manner, a pupil whose score is 63 is said to be one standard deviation below the mean, and the score is written $- 1\sigma$. Such scores are called *standard scores* or *Z-scores.*[7] The T-score, sometimes used as a derived score, is also based upon this idea. The mean is represented as a T-score of 50, and σ as 10. A score $.5\sigma$ below the mean ($-.5\sigma$) would, therefore, be 45, and one $.5\sigma$ above the mean ($+.5\sigma$) would be 55, and so on for other T-scores.[8]

Point scores on tests are also transmuted into letter grades by a similar system. Any pupil who is $+1.5\sigma$ or higher in the distribution is considered *A,* one who is between $+.5\sigma$ and $+1.5\sigma$ is considered *B,* one who is between $-.5\sigma$ and $+.5\sigma$ is considered *C,* one who is between -1.5σ and $-.5\sigma$ is considered *D,* and one who is below -1.5σ is considered *E.* In popular language, a *C* pupil is described as an "average" pupil and an *A* pupil is described as an "outstanding" pupil. This system merely gives definiteness to these descriptions by indicating *how close to the average* (mean) the *C* pupil is, and *how far above the mean* the *A* pupil stands. An *E* pupil is equally "outstanding"; the only difference is that he stands out at the other end of the curve.[9]

Which measure of variability is best? As a rule, σ is regarded as the best measure of variability; and the range is undoubtedly the poorest. The range is subject to all the limitations which the mode has as a measure of average. Just as the mean is greatly influenced by extreme scores, so is σ. Whenever it is desirable, therefore, to avoid the influence of extreme scores, the median is employed as a measure of average, and with it *Q* as a measure of variability. In like manner, when the mean is used as a measure of average, σ is used with it as a measure of variability.

E. Measures of Relationship

The coefficient of correlation. Frequently two or more series of test scores or other quantitative data are available for the same individuals as in Table 19. In such situations it may be important to inquire into the relationship among these measures. What, for example, is the relationship between the chronological age (CA) and the educational age (EA), or between EA and MA in a particular class, or the relationship between IQ and school marks, or the relationship between school marks and amount of study done?

---

[7] For a fuller discussion of Z-scores, see page 306.

[8] For a fuller discussion of T-scores, see pages 306–307.

[9] For a fuller discussion of this procedure, see pages 413–421.

These and similar questions may best be answered by the method of correlation; that is, by obtaining the *coefficient of correlation,* which is a numerical expression of the amount and direction of the relationship between the two series of measures.

Computing the coefficient of correlation by the rank-difference method. The simplest method of determining the correlation is based upon the differences in the rank orders of the two series of measures. Two procedures have been suggested, but the better one is the *rho* (ρ) method devised by Charles E. Spearman. The formula for the Spearman rank-difference method is

$$\rho = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)} .$$

Table 25 illustrates the computation of the coefficient of correlation by the *rho* method, using the EA and MA data in Table 19. The process involves four simple steps, as follows:

Step 1. *Obtain the rank order of all scores in each series.* It is usually desirable, but not necessary, to have one series in order of size. Educational age scores are so arranged in Table 19. Where two or more scores in a series are of the same size, they are assigned their average rank in the series. For example, there are two 183's in X column, and, as there are three scores higher than 183, these two scores are, therefore, assigned the average of the next two ranks, 4 and 5, which is 4.5. In like manner there are four 176's, which receive the average of ranks 10, 11, 12, 13, or 11.5, and two 167's, which receive the average of ranks 18 and 19, or 18.5. The mental age of each pupil is placed in the Y column opposite his educational age in the X column, and the ranks are assigned in the same way. The highest mental age, given a rank of 1, is that of the pupil who ranks 2 in educational age. The pupil who ranks 2 in mental age is 1 in educational age, and the next pupil has a rank of 3 in both columns. It will be noted that there are two 185's, which receive the average of ranks 8 and 9, or 8.5; two 180's, which receive the average of ranks 11 and 12, or 11.5; and three 165's, which receive the average of ranks 16, 17, 18, or 17.

Step 2. *Obtain the differences in rank in the two series.* This column is headed D. The first pupil, with a rank of 1 in educational age and 2 in mental age, is given a difference of $+1$, while the second pupil, with ranks reversed, has a difference of -1. The only advantage in keeping the signs before the differences, D's, is that it affords one check on the accuracy of the work; for the sum of the $+D$'s and the sum of the $-D$'s should be the same. In this example each sum is 33.

Step 3. *Obtain the squares of the differences in ranks.* This is the D^2 column, whose sum is 445, that is, $\Sigma D^2 = 445$.

Step 4. *Substitute in the formula for the value of* ρ. In this case,

$$\rho = 1 - \frac{6 \times 445}{20\,(20^2 - 1)} = 1 - \frac{2{,}670}{7{,}980} = .67.$$

The *rho* (ρ) method has certain definite advantages. It is simple and economical of time if the number of cases is small, possibly not more than 30; and it is especially appropriate if the original data are in ranks. The method also has certain disadvantages. It is time consuming if there is a large number of cases, and it is only an approximate measure of the relationship involved, since it takes into account only the *rank orders* of the differences, rather than their exact magnitudes. Most of the needs of educational measurement require a more exact measure of relationship.

Table 25

COMPUTING THE COEFFICIENT OF CORRELATION BY THE SPEARMAN RHO METHOD

Scores		Ranks		Differences in Ranks	
Educational Age X	*Mental Age Y*	*X*	*Y*	*D*	*D*²
188	208	1	2	+ 1	1
186	218	2	1	− 1	1
185	201	3	3	0	0
183	185	4.5	8.5	+ 4	16
183	165	4.5	17	+12.5	156.25
182	191	6	6	0	0
181	185	7	8.5	+ 1.5	2.25
180	193	8	5	− 3	9
179	181	9	10	+ 1	1
176	165	11.5	17	+ 5.5	30.25
176	187	11.5	7	− 4.5	20.25
176	176	11.5	14	+ 2.5	6.25
176	180	11.5	11.5	0	0
175	166	14	15	+ 1	1
174	197	15	4	−11	121
173	154	16	20	+ 4	16
171	180	17	11.5	− 5.5	30.25
167	165	18.5	17	− 1.5	2.25
167	177	18.5	13	− 5.5	30.25
165	164	20	19	− 1	1

$N = 20$ $\Sigma D^2 = 445$

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)} = 1 - \frac{6 \times 445}{20(20^2-1)} = 1 - \frac{2{,}670}{7{,}980} = 1 - .33 = .67.$$

Computing the coefficient of correlation by the product-moment method. Karl Pearson advocates a method based upon the deviations from the means of each series. This is the product-moment method and gives the so-called Pearson r. The formula is

$$r = \frac{\frac{\Sigma xy}{N} - c_x c_y}{\sigma_x \sigma_y}.$$

The product-moment method is illustrated in Table 26, using the same data as in the rank-difference method. The product-moment method involves four steps, as follows:

Step 1. *For each series obtain the deviation of each score from its mean.* The true mean may be used, but as the true mean is rarely a whole number, it is usually more convenient to work from an assumed mean. Here the assumed mean of X is taken at 176, and that of Y at 185. Each deviation of X is represented by x, and each deviation of Y is represented by y. The signs of the deviations must be indicated.

Step 2. *Obtain the squares of each deviation.* These columns are headed x^2 and y^2.

Step 3. *Obtain the product of each pair of deviations.* That is, each x is multiplied by its corresponding y. The first five products are as follows: $12 \times 23 = 276$; $10 \times 33 = 330$; $9 \times 16 = 144$; $7 \times 0 = 0$; $7 \times -20 = -140$. The others are obtained in a similar manner.

Step 4. *Substitute in the formula.* $\Sigma xy/N$ is 1,262, the algebraic sum of the xy column, divided by N, which is 20. This gives 63.1. c_x is the algebraic sum of the x column divided by N, and c_y is the algebraic sum of the y column divided by N. In this illustration c_x is $+ 1.15$, and c_y is $- 3.1$. Their product, c_xc_y, is $1.15 \times - 3.1 = - 3.565$. But as the formula is $\Sigma xy/N - c_xc_y$, the substitution is $63.1 - (-3.565)$, or $63.1 + 3.565$. σ_x and σ_y are obtained by the usual formula. That is:

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N} - c_x^2} = \sqrt{\frac{831}{20} - (1.15)^2} = \sqrt{41.55 - 1.32} = 6.3,$$

and

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N} - c_y^2} \quad \sqrt{\frac{5{,}296}{20} - (-3.1)^2} = \sqrt{264.8 - 9.61} = 16.0.$$

The product of $\sigma_x\sigma_y$ is $6.3 \times 16.0 = 100.8$.
$r = .66$.

It will be observed that the product-moment method in this situation gives almost exactly the same value as the rank-difference method. This is not a mere coincidence. The differences between the values of r and ρ are likely to be less than .02, although rarely ever exactly the same.

In the illustration below, the data are ungrouped, since only 20 cases are involved. Of course, in the rank-difference method the scores must always be ungrouped, but this is not true when the product-moment method is employed. As a rule, when 40 or more cases are involved, considerable time will be saved in obtaining r by employing a somewhat different procedure, applicable to grouped scores, although the formula is exactly the same. The reader who is interested in this procedure for grouped data is referred to any standard text in educational statistics, of which several are listed in the bibliography at the end of this chapter.

TABLE 26

COMPUTING THE COEFFICIENT OF CORRELATION BY THE PEARSON r METHOD

SCORES		DEVIATIONS		DEVIATIONS SQUARED		PRODUCTS OF DEVIATIONS
X	Y	x	y	x^2	y^2	xy
188	208	+12	+23	144	529	276
186	218	+10	+33	100	1,089	330
185	201	+ 9	+16	81	256	144
183	185	+ 7	0	49	0	0
183	165	+ 7	—20	49	400	—140
182	191	+ 6	+ 6	36	36	36
181	185	+ 5	0	25	0	0
180	193	+ 4	+ 8	16	64	32
179	181	+ 3	— 4	9	16	— 12
176	165	0	—20	0	400	0
176	187	0	+ 2	0	4	0
176	176	0	— 9	0	81	0
176	180	0	— 5	0	25	0
175	166	— 1	—19	1	361	19
174	197	— 2	+12	4	144	— 24
173	154	— 3	—31	9	961	93
171	180	— 5	— 5	25	25	25
167	165	— 9	—20	81	400	180
167	177	— 9	— 8	81	64	72
165	164	—11	—21	121	441	231
$M'_x = 176$	$M'_y = 185$	+63	+100	20)831	20)5,296	+1,438
($N =$ 20)		—40	—162	41.55	264.8	— 176
		20)+23	20)—62	—1.32	—9.61	20)+1,262

$c_x = +1.15$ $c_y = -3.1$ $\sigma^2_x = 40.23$ $\sigma^2_y = 255.19$ $\frac{\Sigma xy}{N} = 63.1$

$c^2_x = +1\ 32$ $c^2_y =$ 9.61 $\sigma_x =$ 6.3 $\sigma_y =$ 16.0.

$$r = \frac{\frac{\Sigma xy}{N} - c_x\ c_y}{\sigma_x\ \sigma_y} = \frac{63.1 - (1.15 \times -3.1)}{6.3 \times 16.0} = \frac{63.1 + 3.565}{100.8} = .66.$$

Interpreting the coefficient of correlation. In interpreting the coefficient of correlation, two things must be considered. The first is the *sign* of the coefficient. The sign indicates the *direction* of the relationship. Positive coefficients indicate direct relationship; that is, there is a tendency for the two series of values to vary in the same direction, high values in one column being associated with high values in the other column, low values in one column being associated with low values in the other column, and so on. On the other hand, negative coefficients indicate inverse relationship; that is, there is a tendency for the two series of values to vary in opposite directions, high values in one column being associated with low values in the other column, and high values in that column being associated with low values in the first column. The coefficient of correlation between EA and MA is +.66 in the foregoing illustration.[10] This means that there is a tendency for pupils whose EA's are high to have high MA's, and conversely. A glance at the data in Table 25 will reveal that such is the case. If the reader will examine the first two columns in Table 18, however, he will observe a different relationship. Here it will be apparent that there is a tendency for pupils above average in educational age to be below average in chronological age, while those who are below average in educational age are above average in chronological age. This inverse relationship between EA and CA is a matter of common observation in almost every classroom. The overage pupils in any grade are generally well toward the foot of the class educationally, whereas the underage pupils are generally well toward the head of the class. In this instance the Pearson r between CA and EA is $-.37$, while ρ is slightly less, or $-.34$. The first thing to consider, then, is the sign of the coefficient, for that indicates the direction of the relationship.

Another thing is equally important and far more difficult to interpret; that is, the *magnitude* or *size* of the coefficient. The *size* of the coefficient indicates the *degree* or closeness of the relationship, just as the *sign* of the coefficient indicates the *direction* of the relationship. The minimum coefficient is .00, which indicates no consistent relationship whatsoever. From this minimum value the coefficients increase in both directions until +1.00 is reached for one limit, and -1.00 for the other. It should be noted that both +1.00 and -1.00 indicate equally close relationship, for both are perfect. Their one important difference is in direction, the former being direct and the latter being inverse. In like manner, all other values of the same size, such as +.50 and $-.50$, indicate equally close relationship. It is the size, and not the sign, of the coefficient that gives the clue to the closeness or degree of relationship.

The problem, then, is to know how close a relationship is indicated by a coefficient of correlation of a given magnitude, regardless of sign. For example, how close a relationship is indicated by a coefficient of .60? Unfortunately, there is no simple way of answering such a question. Attempts to indicate this relationship by some descriptive adjective, such as "high" or "marked," are vague and often misleading, to say the least. As a matter of fact, a coefficient of .60 might be regarded as high for one type of situation and low for another. For example, a coefficient of .60 between a general intelligence test administered at the beginning of the year and school marks recorded at the end of the year might be regarded as high, because such coefficients usually fall well below that. But a coefficient of .60 between scores on two forms of this intelligence test administered the same day, or between scores on one form administered at the beginning of the year and scores on another form of the same test administered at the end of the year, would be unusually low. In other words, "high" and "low" have only *relative* meaning. Before an interpretation can be made of a coefficient on this basis, the reader must at least know what the central tendency of such coefficients for similar data is; and a knowledge of the total distribution of the coefficients is desirable. Even then this method does not tell us just how close the relationship is. At best, the answer to the questions is vague and indefinite.

---

[10] In actual practice the + is usually omitted in printing. If no sign appears before the coefficient, it is always understood to be positive.

Of the various attempts to give a definite mathematical interpretation of the degree of relationship implied by a coefficient of correlation of a given magnitude, only one will be described, and that briefly. For this purpose Kelley has proposed the *coefficient of alienation,* designated by the letter k. This is a measure of the departure from perfect agreement or correlation. The formula is $k=\sqrt{1-r^2}$. Take as an illustration the .60 referred to above. Substituting in the formula, we have

$$k = \sqrt{1 - .60^2} = \sqrt{1 - .36} = \sqrt{.64} = .80.$$

When r is .60, k is .80, which means that the departure from perfect agreement is 80 per cent. Stated the other way around, when there is a coefficient of correlation of .60 between two things, it is possible to predict the values in one series from those in the other series only 20 per cent better than chance. It may be disquieting to discover that even with an r of .80, k is .60, which means that perfect agreement is only 40 per cent better than chance. Taylor and Russell,[11] for example, point out that for some purposes the formula produces "unwarranted pessimism." It is important to note that while the departure from perfect agreement is large with coefficients of the size usually found for educational data, the majority of the errors involved are not large. For example, although the departure from perfect agreement when r is .60 is 80 per cent, a table which Otis[12] gives indicates that the disagreement that exceeds two thirds of a standard deviation is only about 26 per cent.[13] In like manner, an r of .80, with a 60 per cent departure from perfect agreement, has less than 11 per cent of disagreement that exceeds two thirds of a standard deviation.

---

[11] H. C. Taylor and J. T. Russell, "The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables," *Journal of Applied Psychology,* 23: 565–578, October, 1939.

Uses of the coefficient of correlation. One of the most important uses of the coefficient of correlation is for determining the validity of a test. It will be recalled that there are two types of validity, or rather two methods of judging the validity of a test: namely, curricular and statistical. The former is subjective, and the latter is objective. Curricular validity is determined by examining the content of the test itself and judging the degree to which it is a true measure of the important objectives of the course, or a truly representative sampling of the essential materials of instruction. Statistical validity is determined by setting up a criterion of the thing which it is desired to measure, and then computing the coefficient of correlation between the test scores and the criterion. The product-moment method is generally used. The r so obtained is called a *validity coefficient,* and is interpreted like any other coefficient of correlation.

A second use of the coefficient of correlation is for determining the reliability of a test. Since reliability is the degree of consistency with which the test measures whatever it does measure, a convenient way to determine reliability is by computing the coefficient between two forms of the same test, two halves of a test, or two applications of the same test. The product-moment method is generally employed for this purpose.

When the scores on one half of the test are correlated with the scores on the other half, the reliability of the half-test is, of course, obtained. From this coefficient the reliability of the entire test can be estimated by the Spearman-Brown prophecy formula. This formula for the estimated coefficient, r_{nn}, is

$$r_{nn} = \frac{nr_{1I}}{1 + (n-1)\,r_{1I}}\cdot$$

In this formula n is the number of times the test whose reliability is to be estimated is longer than the one whose reliability is known. When the reliability of the whole test is being estimated from the half-test, the value of n is 2. This means that the desired reliability coefficient is obtained by a simple substitution, as follows:

---

[12] Arthur S. Otis, *Statistical Method in Educational Measurement,* page 225. Yonkers: World Book Company, 1925.

[13] Otis expresses the disagreement in terms of median deviations, each of which is 68σ, or approximately two thirds of a standard deviation.

$$\text{Estimated } r = \frac{2 \times r \text{ of half-test}}{1 + r \text{ of half-test}}.$$

For example, suppose the correlation of the even-numbered items with the odd-numbered items on a given test is .60, the estimated r for the whole test is

$$\frac{2 \times .60}{1 + .60} = \frac{1.20}{1.60} = .75.$$

The same formula can be used for estimating the reliability of the test if increased to any required length. Since .75 is still rather low for a reliability coefficient, the teacher might wish to estimate the reliability of the test if it were further increased in length. The effect of doubling the length of the test, for example, would be estimated by substituting in the same manner:

$$\frac{2 \times .75}{1 + .75} = \frac{1.50}{1.75} = .86.$$

A third important use of the coefficient of correlation in educational measurement is for prediction. All sound guidance is conditioned upon the ability to foresee, or predict, the future. For this purpose, the coefficient of correlation has shown itself a valuable tool. As an illustration, one might ask which of the following three measures available at the beginning of the year affords the best basis for judging the probable success of pupils in plane geometry: (1) achievement in algebra, (2) intelligence test scores, or (3) scores on a geometry aptitude test? The answer is, whichever has shown the highest correlation with geometry achievement. For example, if it has been found that algebra and geometry marks give a coefficient of .40, intelligence scores and geometry marks give a coefficient of .50, and geometry aptitude scores and geometry marks give a coefficient of .60, then it is evident that the poorest basis of prediction is algebra marks, and the best is scores on a geometry aptitude test.

Sometimes it is desirable to know what the correlation between two things would be if some other factor or factors were "held constant." For this purpose, what is known as *partial correlation* is used. Again, it is sometimes desirable to know what the correlation would be between one thing, such as geometry marks, and a combination of two or more things, such as geometry aptitude scores, intelligence scores, and algebra marks. For this purpose, what is known as *multiple correlation* is used. The first step in both partial and multiple correlation is to obtain the simple correlations, or Pearson *r's*, for all the variables involved. The actual predictions for an individual pupil, regardless of whether simple or multiple correlation is used, necessitate substituting in a *regression equation.* As the above techniques are not commonly employed by the average teacher, and are therefore beyond the scope of this book, they will not be illustrated.

F. Measures of Error

Errors in educational measurement may be conveniently grouped into three types, according to source:

1. Errors of technique:
   a. Arithmetical errors in computation, etc.
   b. Use of inappropriate measures.
2. Errors of measurement:
   a. Imperfect measuring instruments.
   b. Lack of skill in the measurer.
   c. Fluctuations in the persons measured.
3. Errors of sampling:
   a. Selection or bias in sampling.
   b. Chance fluctuations in random sampling.

Errors of technique. Obvious types of errors are mistakes in adding scores and various computational errors in statistical analysis. The only protection against such errors is the exercise of great care. Likely to be more serious are the errors due to the use of inappropriate measures for the data in hand. It is poor technique to introduce more refined measures than the data warrant or the purpose requires. All statistical formulas are based upon certain assumptions which are often not fully met in actual practice. The following are common examples: Computations based upon data in frequency distribution assume that the scores are uniformly distributed within the several intervals or that the mid-point of each interval may be used to represent the average value of all scores in the interval. The Pearson r assumes linearity of relationship among the data. Most formulas are based on the assumption of a normal distribution of the measures. Whenever the data in a given situation fail to conform to these assumptions, certain errors are introduced. Fortunately, in actual practice, these errors are often not great enough to introduce serious errors of interpretation. But gross errors due to the use of inappropriate techniques are sufficiently numerous to warrant extreme caution. Furfey and Daly [14] made a study of the articles in recent issues of five professional journals using the product-moment correlation, and came to the conclusion that this technique is employed "with little regard to the fulfillment of the necessary antecedent conditions." In fact, in 60 of the 63 articles studied they found that "their authors have left themselves open to the suspicion of having employed the correlation technique in a way which is meaningless, if not positively misleading."

---

[14] Paul Hanly Furfey and Joseph F. Daly, "Product-moment Correlation as a Research Technique in Education," *Journal of Educational Psychology,* 26: 206–211, March, 1935.

Errors of measurement. There are three possible sources of errors in measurement, even when there are no computational errors and when the most appropriate statistical analysis has been employed. In the first place, no measuring instrument is perfectly valid or perfectly reliable. In the second place, the personal equation of the examiner must be reckoned with. Inexperienced examiners may allow too much or too little time in administering the test, or may otherwise depart from standardized procedure in administering the test or in scoring the papers. In the third place, there is likely to be great variability in the responses of the subjects taking the test. Accidental occurrences, such as the breaking of a pencil on timed tests, fluctuations in motivation, fatigue, and other physical and mental factors may seriously affect the test results.

It will be noted that some errors of measurement are systematic and tend to affect all individuals alike. Allowing too much or too little time on a test of reading speed is an example. On the other hand, many errors of a variable character occur affecting the individuals unequally or in different directions. Sensory defects, health conditions, and motivation are examples of conditions that produce variable errors in measurement. The effects of these errors are briefly presented in table form.

MEASURE	CONSTANT ERRORS	VARIABLE ERRORS
Central Tendency	Increased or decreased by amount of the error	Usually tend to offset or balance each other
Variability	Little or no effect	Usually made too large
Relationship	Little or no effect	Usually made too small

It will be observed that constant errors affect most seriously measures of central tendency, and unfortunately there are no formulas for this correction. Such formulas as exist usually take into account variable errors of reliability only. Appropriate formulas for errors of validity do not exist.

Errors of sampling. It is usually impractical to measure all the cases of a given type. For example, it would be a formidable task to obtain the IQ's of all high-school freshmen in a state, or the difficulty of each word in a series of textbooks. Fortunately, it is not necessary to do so. It has been found possible to estimate the range of errors within which the true measure may be expected to lie. But to do so, it is necessary to have a representative sampling of the total population. Against errors in a selected or "hand-picked" sampling there is no statistical protection. An adequate sampling must be chosen in a random manner; and the larger the sampling, the better, although increasing the number of cases does not in itself eliminate the possibility of error. The formula for the standard error of the mean is

$$\sigma_M = \frac{\sigma}{\sqrt{N}}.$$

The formula for the standard error of the standard deviation is

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}}.$$

The corresponding formula for the coefficient of correlation is

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}.$$

One of the most useful formulas of all is that for the error of a difference between means of two samples. This formula is

$$\sigma_{M_1 - M_2} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2} - 2 r_{12} \sigma_{M_1} \sigma_{M_2}}.$$

If the two series of measures are uncorrelated, the formula is shortened to

$$\sigma_{M_1 - M_2} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}}.$$

Attention should perhaps be called to the fact that *N* is in the denominator of these formulas. This means that as *N* increases the errors decrease.

Errors are often expressed in terms of the probable error, or P.E., instead of the standard error, or σ. The only difference between these measures is in length, the former being approximately two thirds the latter. The formula in all cases is $\text{P.E.} = .6745\sigma$.

The above formulas do not take into account systematic errors of any type or any errors of validity. They merely make allowance for chance errors in sampling and in the reliability of measurement.

The ordinary formulas considered here are based upon certain assumptions that may not always be true for the data under consideration. If the sampling used is a truly random one and involves as many as 30 or 40 cases, the distribution of errors is likely to be approximately normal, however, an assumption on which the usual formulas are based. When the sample contains fewer than 30 cases, even if random, its σ is likely to be smaller than the σ of the total population. In such situations the so-called *small sample theory,* which employs $N - 1$ or some other correction instead of N, will give more trustworthy results. Readers interested in this theory will find the discussion by Lindquist[15] is among the most satisfactory available.

Interpretation of measures of error. In all cases the size of the standard or probable error gives the probability that the true value for the total population lies within various ranges above and below the obtained values for the sample. For example, $r = .60 \pm .05$ tells us that a certain obtained coefficient of correlation has a probable error of .05. The probabilities that the true value for the entire population lies within various ranges determined by various multiples of the P.E. are as follows:

1 to 1 for a range of -1 P.E. to $+1$ P.E.;
4.6 to 1 for a range of -2 P.E. to $+2$ P.E.;
22 to 1 for a range of -3 P.E. to $+3$ P.E.; and
142 to 1 for a range of -4 P.E. to $+4$ P.E.

In this case where r is .60 and P.E. is .05 the chances are

1 to 1 that the true r is between .55 and .65;
4.6 to 1 that the true r is between .50 and .70;
22 to 1 that the true r is between .45 and .75; and
142 to 1 that the true r is between .40 and .80.

It can be seen that the larger the errors the wider the expected range. *It is important to remember that statistical measures can rarely be taken at their face value.*[16] It is useful to regard a pupil's score on a test as a *range,* rather than as a point.

---

[15] E. F. Lindquist, *Statistical Analysis in Educational Research,* Chapter III. Boston: Houghton Mifflin Company, 1940.

[16] Walter S. Monroe and Max D. Engelhart, *The Scientific Study of Educational Problems,* page 158. New York: The Macmillan Company, 1936.

The interpretation of a difference between means is commonly expressed in terms of the ratio of the difference to its P.E. This is sometimes called the *critical ratio.* If this ratio is 4 or more, the difference is usually regarded as *statistically significant.* This means that the difference is probably real and not merely chance, but in and of itself it does not tell whether the difference is of practical significance. Table 27 gives the chances in 100 that a difference between means is significant for ratios of various magnitudes.

TABLE 27

THE CHANCES IN 100 THAT A TRUE DIFFERENCE EXISTS BETWEEN TWO MEANS

$\frac{M_1 - M_2}{\text{P.E.}_{M_1 - M_2}}$	CHANCES IN 100 OF A TRUE DIFFERENCE
.00	50
.20	55
.40	61
.60	66
.80	71
1.00	75
1.20	79
1.40	83
1.60	86
1.80	89
2.00	91
2.20	93
2.40	95
2.60	96
2.80	97
3.00	98
3.20	98.5
3.40	98.9
3.60	99
3.80	99.5
4.00	99.7

SELECTED REFERENCES FOR FURTHER READING

Edwards, Allen L., *Statistical Analysis for Students in Psychology and Education.* New York: Rinehart & Company, Inc., 1946. 360 pages.

Garrett, Henry E., *Statistics in Psychology and Education* (Second Edition). New York: Longmans, Green & Company, 1937. 493 pages.

Gray, Clarence T., and Votaw, David F., *Statistics Applied to Education and Psychology.* New York: The Ronald Press Company, 1939. 278 pages.

Guilford, J. P., *Fundamental Statistics in Psychology and Education.* New York: McGraw-Hill Book Company, Inc., 1942. 333 pages.

Harper, F. H., *Elements of Practical Statistics.* New York: The Macmillan Company, 1930. 324 pages.

Holzinger, Karl J., *Statistical Methods for Students in Education.* Boston: Ginn and Company, 1928. 372 pages.

Lindquist, E. F., *A First Course in Statistics.* Boston: Houghton Mifflin Company, 1938. 226 pages.

Monroe, Walter S., and Engelhart, Max D., *The Scientific Study of Educational Problems.* New York: The Macmillan Company, 1936. 504 pages.

Odell, C. W., *An Introduction to Educational Statistics.* New York: Prentice-Hall, Inc., 1946. 269 pages.

Peters, Charles C., and Van Voorhis, Walter R., *Statistical Procedures and Their Mathematical Bases.* New York: McGraw-Hill Book Company, Inc., 1940. 516 pages.

Sorenson, Herbert, *Statistics for Students of Psychology and Education.* New York: McGraw-Hill Book Company, Inc., 1936. 373 pages.

Walker, Helen M., *Elementary Statistical Methods.* Henry Holt & Co., Inc., 1943. 368 pages.

Van Omer, Edward B., and Williams, Clarence O., *Elementary Statistics for Students of Education and Psychology.* New York: Longmans, Green and Company, 1940. 111 pages.

CHAPTER IX

The Graphical Representation of Educational Data

A. The Value of Graphs

"One picture is worth ten thousand words." So runs an old Chinese proverb. "There is a magic in graphs," says a modern scientific writer.[1] He describes the dynamic role of the graphical representation of numerical data as follows:

> Words have wings, but graphs interpret. Graphs are pure quantity stripped of verbal sham, reduced to dimension, vivid, unescapable. . . . Wherever there are data to record, inferences to draw, or facts to tell, graphs furnish the unrivaled means whose power we are just beginning to realize and to apply.

There can be little doubt that the graphical representation of educational data is a valuable supplement to statistical analysis and summarization. The psychological value of graphs in the testing program may be considered under three headings: They attract attention, they clarify the meaning, and they aid retention.

Graphs attract attention. In the first place, the graph or chart tends to attract the reader's attention. Advertisers employ a wide variety of pictures, charts, and diagrams, for they realize that the first step in making a sale is to attract the prospective customer's attention. They have learned that pictures will do this where numerical data and printed material will not. The average reader is likely to give scant attention to the ordinary printed matter in a school report and be wholly unimpressed by the appalling mass of tabular data often piled up at the end, but his eye is sure to be arrested by any picture or chart that may happen to be included. And this may lead him to read the entire discussion. There is evidence that school administrators are beginning to learn this lesson.[2]

Graphs clarify points. In the second place, the graph is often an effective method of clarifying a point. One small chart will often make a point clearer than a dozen tables or paragraphs. It is

---

[1] Henry D. Hubbard, quoted by W. C. Brinton, *Graphic Presentation,* page 2. New York: Brinton Associates, 1939.

[2] *Cf.* Douglas E. Scates, "Reporting, Summarizing and Supplementing Educational Research," *Review of Educational Research,* 12: 558–574, December, 1942.

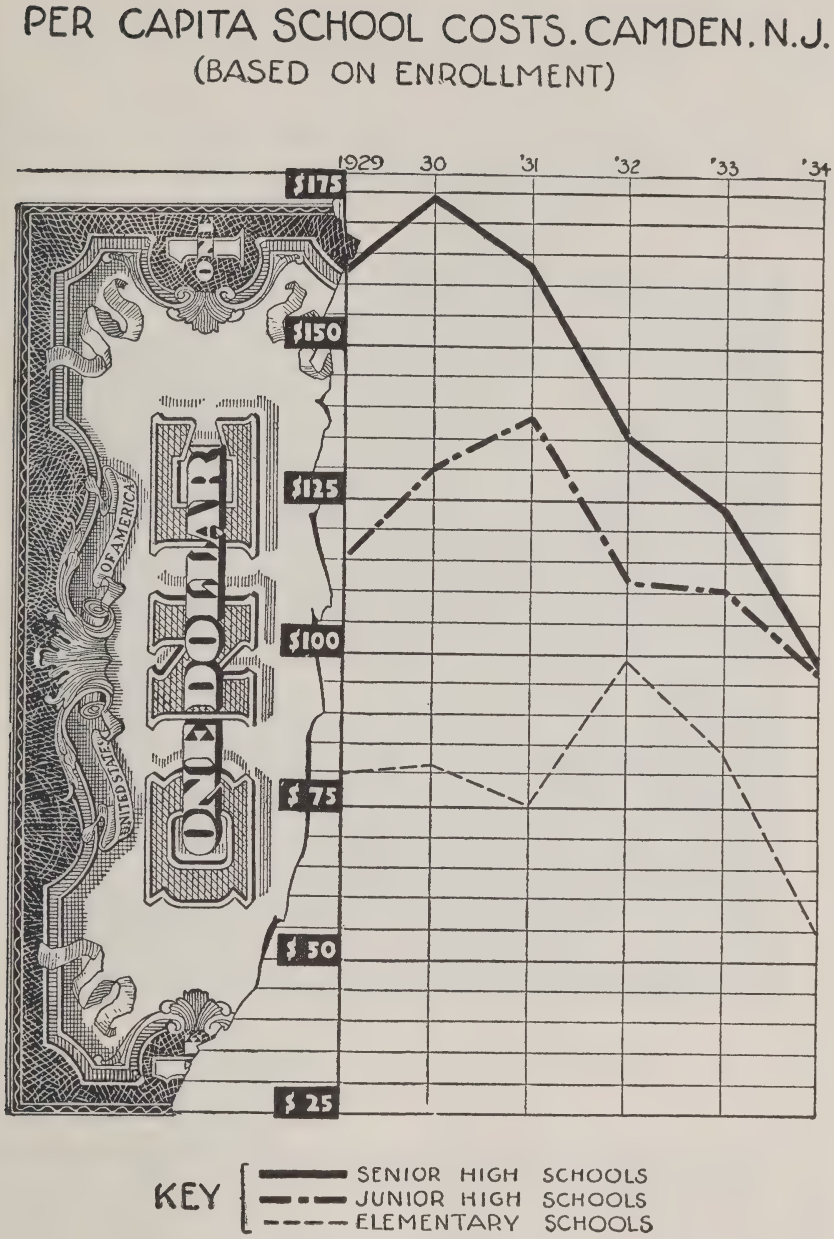


Figure 9. An Effective Chart from a City Superintendent's Annual Report. (From *Reconstruction and the Schools*, the Annual Report of the Camden, New Jersey, Public Schools, 1934, page 59.)

sometimes said that the facts speak for themselves. In reality, statistics often stand speechless and silent, tables are tongue-tied, and only the chart cries aloud its message to all the world. Ordinary numerical data are quite abstract; they convey their meaning vaguely and with effort to the average mind. The picture or graph is a more concrete representation of the matter.

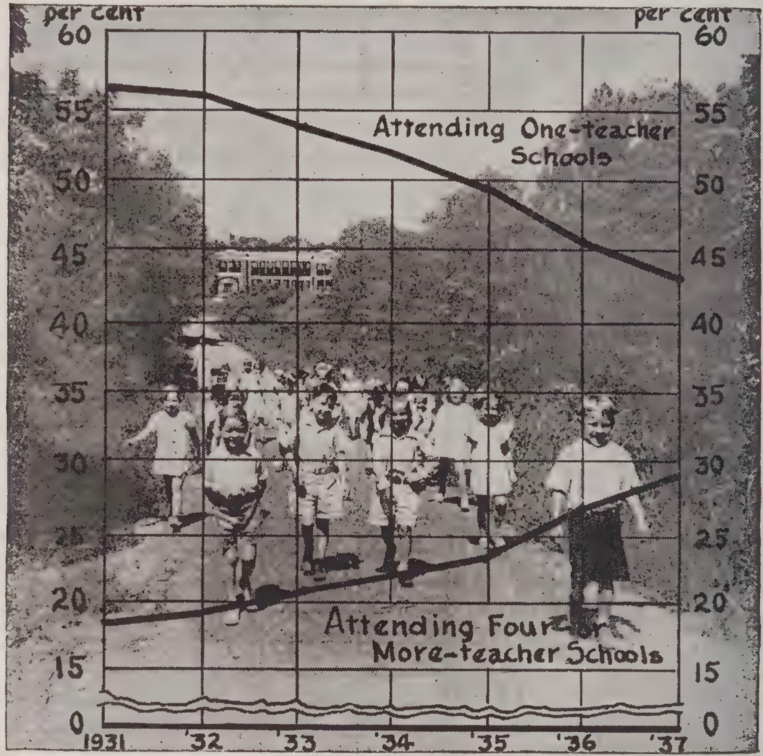
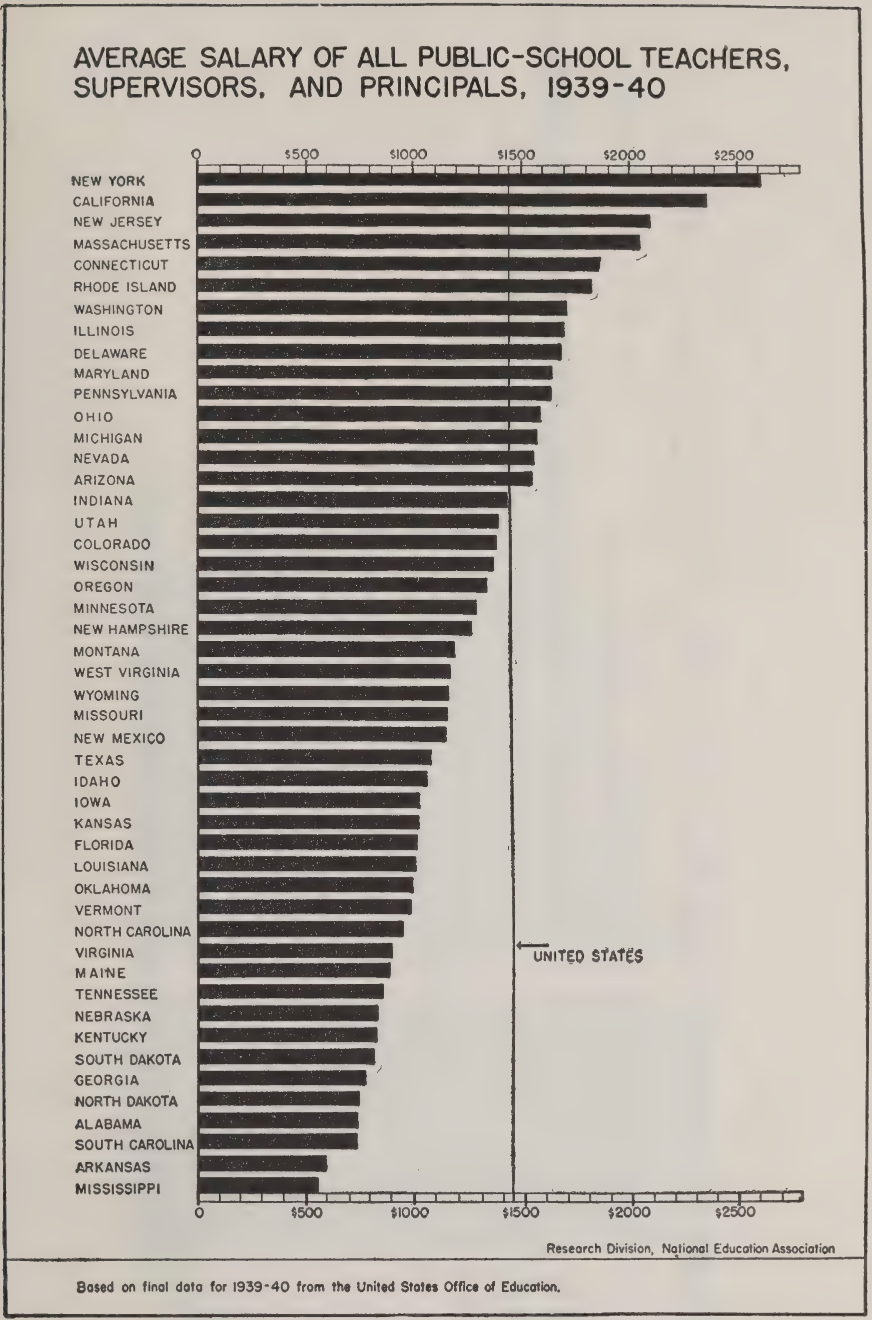


Figure 10. Trends in Elementary-School Enrollments in Kentucky for a Six-Year Period.

School administrators in recent years have been making effective use of the graphical representation of educational data. Figure 9, found in a city superintendent's annual report,[3] is a good example. The dollar bill attracts the reader's attention, and the line graphs show him at a glance the downward trend of school expenditures. Figure 10 shows the trends in elementary school enrollments in

---

[3] *Reconstruction and the Schools,* page 59. Annual Report of the Board of Education, Camden, New Jersey, Public Schools for the School Year Ended June 30, 193



AVERAGE SALARY OF ALL PUBLIC-SCHOOL TEACHERS, SUPERVISORS, AND PRINCIPALS, 1939-40

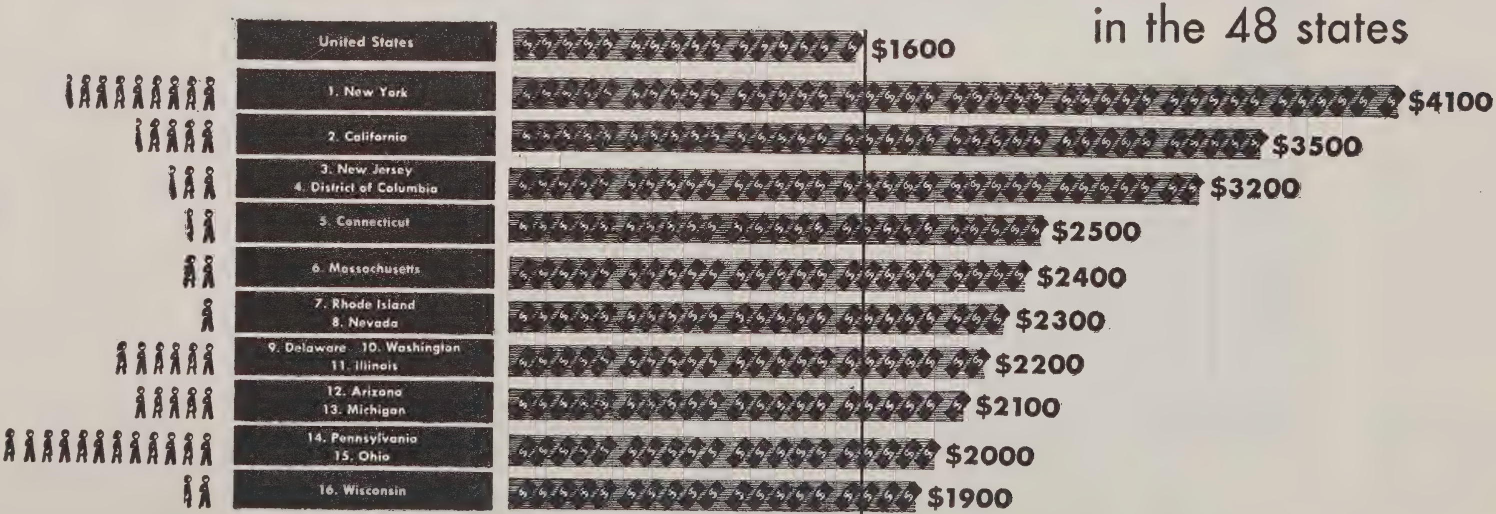
Research Division, National Education Association

Based on final data for 1939-40 from the United States Office of Education.

Figure 11. A Sample Bar Graph Shows Striking Salary Differences Among the 48 States.

Current Expenditure for the Median Classroom Unit

in the 48 states



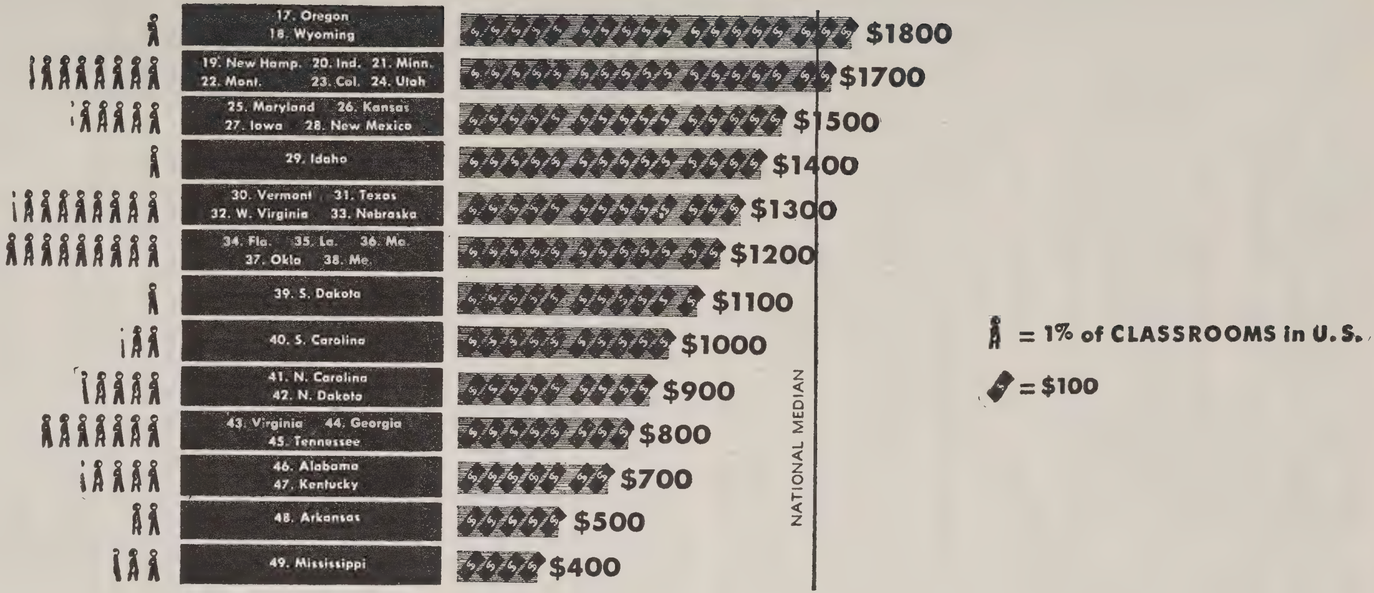


Figure 12. A More Complex Bar Graph with High Attention Value.

Kentucky for a six-year period.[4] The picture of the children on their way from school, as a background, has attention-getting value, and the reader can hardly escape noting the decline in attendance in one-teacher schools along with the corresponding increase in attendance in the larger schools.

Educational research workers have also found that their major findings are most effectively presented in graphical form. Figure 11[5] prepared by the Research Department of the National Education Association and Figure 12 by Norton and Lawler[6] represent strikingly the enormous inequalities among the states in the support of public education.

Graphs aid retention. It has also been found that the graphical presentation of certain types of data is a definite aid to recall. Washburne[7] compared the efficiency of graphical, tabular, and textual modes of presenting historical data to pupils in the junior high school. The material, which dealt with certain specific quantitative facts, was kept constant, but the mode of presentation varied. Sometimes it appeared as a statistical table, sometimes as a bar graph, a pictograph, or a line graph, and at other times it was presented in ordinary paragraph form. Among the conclusions arrived at by the author were the following:[8]

> 1. The paragraph is, in general, the form which is least favorable to recall of quantitative data, whether general or specific.
> 2. The bar graph is the form most favorable to the recall of relative amounts (static comparisons) when the comparisons called for involve a fair degree of difficulty. For very simple data some form of pictograph may be more favorable to the recall of relative amounts than the bar graph.
> 3. The line graph is the form most favorable to the recall of relative increase, decrease, and fluctuation (dynamic comparisons).
> 4. The statistical table is the form most favorable to the recall of specific amounts.

One study[9] on graph interpretation in the elementary schools points out that little is known regarding the comparative value of various graphs, although the circle graph appears to be easiest and the line graph most difficult, with the bar graph occupying an intermediate position. Her results indicated that a mental age of 14 years was required for the satisfactory interpretation of bar and line graphs without specific instruction in reading materials presented in graphical form.

---

[4] Leonard E. Meece and Maurice F. Seay, *Financing Public Elementary and Secondary Education in Kentucky,* page 73. Bulletin of the Bureau of School Service, Vol. XII, No. 1, Lexington: University of Kentucky, 1939.

[5] "Federal Aid for Education, A Review of Pertinent Facts," *National Education Association Research Bulletin,* 20: 129, September, 1942.

[6] John K. Norton and Eugene S. Lawler, *Unfinished Business in American Education,* page 13. Washington: American Council on Education, 1946.

[7] John Noble Washburne, "An Experimental Study of Various Graphic, Tabular and Textual Methods of Presenting Quantitative Material," *Journal of Educational Psychology,* 18: 361–476, September and October, 1927.

[8] *Ibid.,* page 475.

[9] Sister Clara Francis Bamberger, "Interpretation of Graphs at the Elementary School Level," *Educational Research Monographs,* 13: 1–62, May 1, 1942.

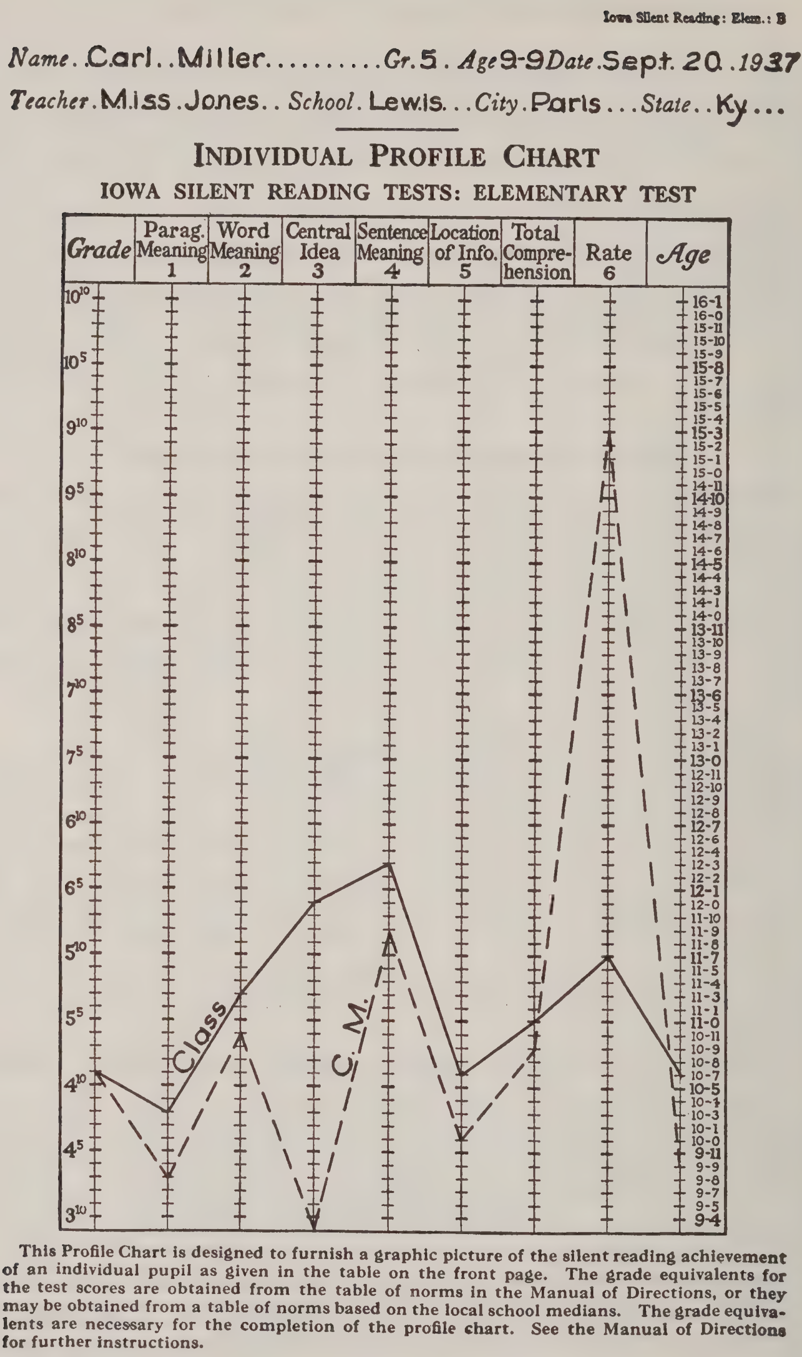
These findings appear to be in line with a principle of learning abundantly supported by experimental evidence: namely, that the method of presentation which makes the meaning clearest is most favorable to learning and recall. It is important to recognize that neither statistical nor graphical methods bestow precision upon data. They are merely useful ways of expressing whatever accuracy exists.

B. Representing the Record of an Individual

There is no more striking way of representing the test record of an individual pupil than by means of a graph. Such a graphical picture of the strong and weak points of a single person is called a *profile*. Sometimes the term *psychograph* is used. Many publishers of standard tests provide blank forms for showing these profiles. Usually they are shown on the first or last page of the test, where they can easily be detached for filing. Some publishers issue these profile blanks as separate cards of a heavier stock.

Profiles of a single subject. Figure 13 shows the profile of a fifth-grade pupil on the Iowa Silent Reading Tests. The profile for the class, based on medians, is also shown. It is apparent that C.M. is weak on Test 1, paragraph meaning, and Test 3, selecting the central idea of a paragraph. The class as a whole is weak on Test 1, paragraph meaning, and on Test 5, locating information, and C.M. is slightly below the class average. His score in reading rate is unusually high. Such a profile enables a teacher to see at a glance the outstanding points of an individual pupil in a given subject.

Profiles for a series of subjects. Profiles are especially useful in representing a pupil's record on two or more subjects. Most test batteries provide a convenient form for such a profile. Figure 14 shows the profile for an eleventh-grade pupil on the Progressive Achievement Tests, Advanced Battery. Some of the irregularities are very striking. It will be noted that the reading vocabulary score is just above the class average and that the reading comprehension score is much below the class average. John Ford is also weak in language, a subject in which the class as a whole made its poorest score. After a period of remedial instruction, the purpose of which is to strengthen the weak points of individual pupils, it is a good practice to give a second form of the same test. A second



Iowa Silent Reading: Elem.: B

Name. Carl. . Miller. . . . . *Gr.* 5 . *Age* 9-9 *Date*. Sept. 20 . 1937

Teacher. Miss . Jones. . *School*. Lewis. . . *City*. Paris . . . *State*. . Ky . . .

INDIVIDUAL PROFILE CHART

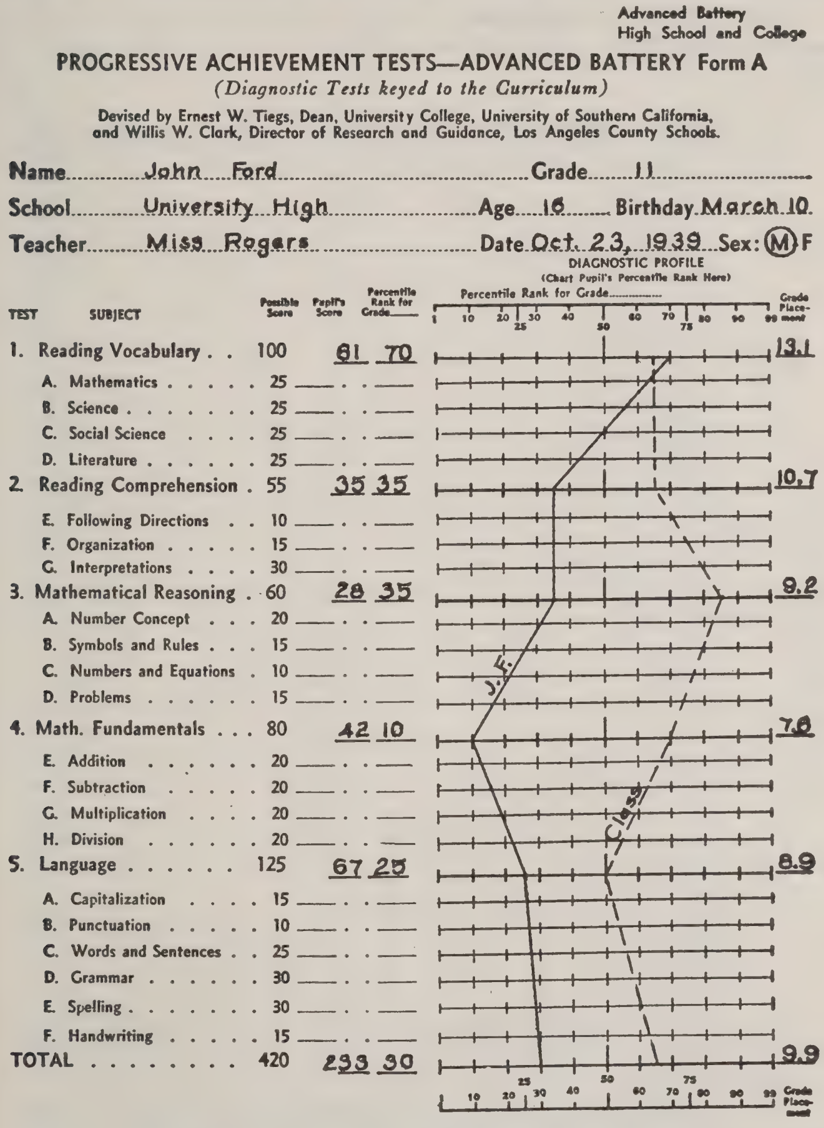
IOWA SILENT READING TESTS: ELEMENTARY TEST

This Profile Chart is designed to furnish a graphic picture of the silent reading achievement of an individual pupil as given in the table on the front page. The grade equivalents for the test scores are obtained from the table of norms in the Manual of Directions, or they may be obtained from a table of norms based on the local school medians. The grade equivalents are necessary for the completion of the profile chart. See the Manual of Directions for further instructions.

Figure 13. Profile of a Pupil and the Fifth-Grade Class of Which He Is a Member. (Published by World Book Company, 1933.)

profile drawn in a different color upon the same sheet is one of the best ways of revealing the progress made.

Profiles showing achievement and intelligence. It is often helpful to represent on a single profile both the achievement of a pupil and his intelligence. Figure 15 shows the profile issued by



Advanced Battery
High School and College

PROGRESSIVE ACHIEVEMENT TESTS—ADVANCED BATTERY Form A

(Diagnostic Tests keyed to the Curriculum)

Devised by Ernest W. Tiegs, Dean, University College, University of Southern California, and Willis W. Clark, Director of Research and Guidance, Los Angeles County Schools.

Name John Ford Grade 11
School University High Age 16 Birthday March 10
Teacher Miss Rogers Date Oct. 23, 1939 Sex: [✓ M] F

TEST	SUBJECT	Possible Score	Pupil's Score	Percentile Rank for Grade	Grade Placement
1.	Reading Vocabulary	100	61	70	13.1
	A. Mathematics	25			
	B. Science	25			
	C. Social Science	25			
	D. Literature	25			
2.	Reading Comprehension	55	35	35	10.7
	E. Following Directions	10			
	F. Organization	15			
	G. Interpretations	30			
3.	Mathematical Reasoning	60	28	35	9.2
	A. Number Concept	20			
	B. Symbols and Rules	15			
	C. Numbers and Equations	10			
	D. Problems	15			
4.	Math. Fundamentals	80	42	10	7.6
	E. Addition	20			
	F. Subtraction	20			
	G. Multiplication	20			
	H. Division	20			
5.	Language	125	67	25	8.9
	A. Capitalization	15			
	B. Punctuation	10			
	C. Words and Sentences	25			
	D. Grammar	30			
	E. Spelling	30			
	F. Handwriting	15			
TOTAL		420	233	30	9.9

Figure 14. The Profile of an Eleventh-Grade Pupil on the Progressive Achievement Test, Advanced Battery. (Published by California Test Bureau, 1937.)

the Educational Test Bureau for representing unit scores in achievement and aptitude. Another method of combining the intelligence score with the achievement record is to draw a red line across the ordinary profile at the level which represents the mental age of the pupil at the time the achievement tests were given.

Use of profiles in guidance. Profiles of individual students have an important place in guidance. Figure 16 is an example of a

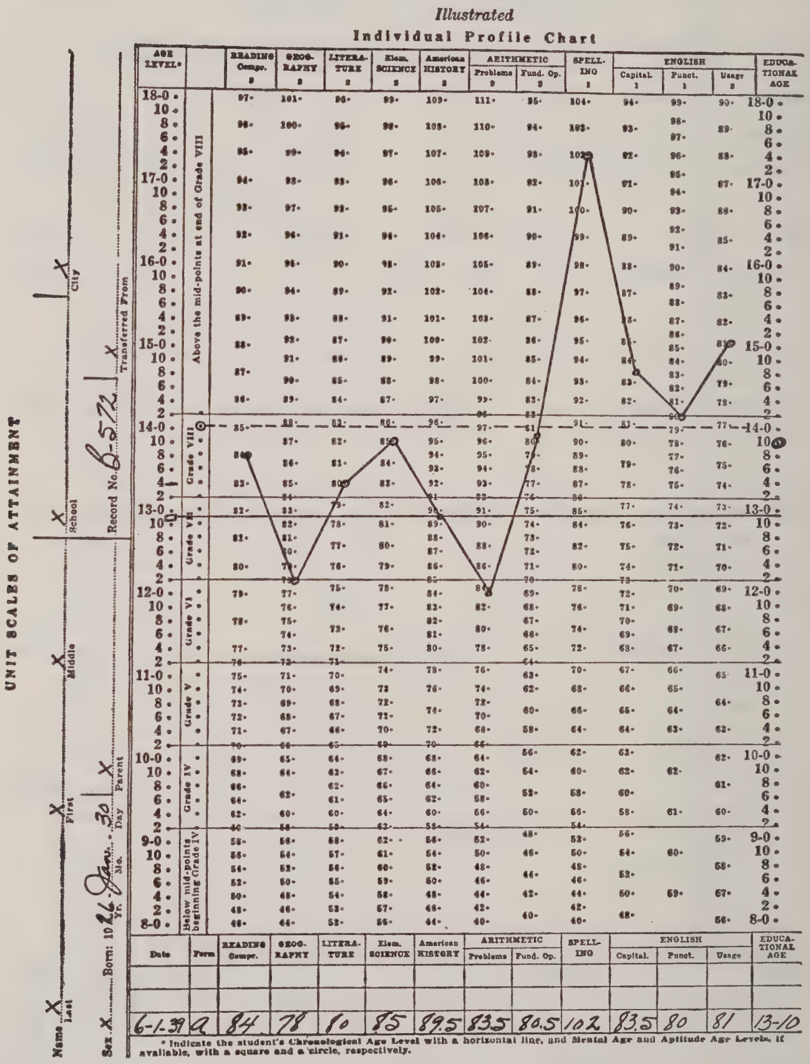


Figure 15. A Profile for Representing Unit Scores in Achievement and Aptitude. (Published by Educational Test Bureau, 1937.)

Illustration 20

PSYCHOLOGICAL TEST PROFILE

Educ. 70---Educ. Measurements.

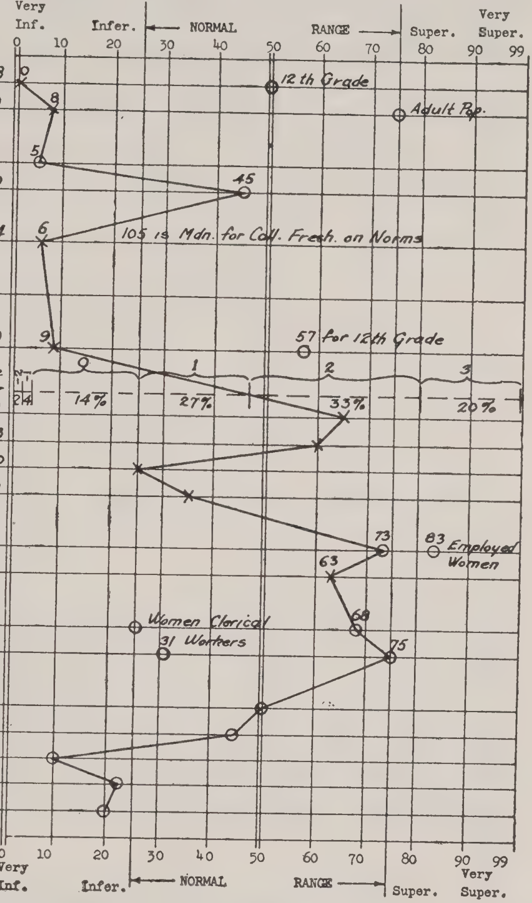
The Pennsylvania State College

Date 3-10-39 Name H. L. J. (1st) (last) Sex Female Age

School Education Class Junior Sem. in College 5th Major Spanish Minor English

Type and Name of Test	Nature of Population	Raw Score
Tests taken during 1st sem. (unless other date)	1938-1939	
GENERAL INTELL.		
1. Henmon-Nelson H.S. Form A	49 Ed. 70 Students	58
2. Otis Q-S; Gamma Form A	57 Ed. 70 Students	49
READING Iowa Silent Adv A		
1. Rate	End Grade 13	18
2. Comprehension		140
VOCABULARY		
1. Inglis-A	156 Liberal Arts Women	94
EIGHTH GRADE ACH.		
1. Batting No	record	
H. S. ACHIEVEMENT Test		
1. Myers-Ruch-A	54 Ed. 70 Students	50
SUBJECT ACHIEVEMENT College Courses	Present Grade dist. of Coll. →	
1. Eng. Lit.	4 Courses	2.0
2. Spanish (2 Best)	6 Courses	1.83
3. Eng. Comp. 2	2 Courses	0.50
4. Soc. Sciences (2 Lowest)	4 Courses	1.00
MECH. APTITUDE		
1. Minn. Paper	247 Liberal Arts Fresh.	43
2. Form Board AA	45 Ed. 70 Students	"
CLERICAL APT. Minn Clerical:		
1. Numbers	232 Employed	124
2. Names	Adult Woman	137
PERSON. ADJUST.		
1. Bell Invent. Home	243 College Women	7
2. Health	"	6
3. Social	"	29
4. Emotional	"	17
5. Total Adj.	"	59

Percentile Scores (% of a certain population exceeded). 0 = population of test norms: X = special group.



COMMENTS:

Got a 0 in College Algebra that follows 1 entrance unit; also got 0 in General Biol., lecture course.

Send to Reading Clinic for vision & reading tests. Why rate of reading so low? Send to Psychological Clinic to take Strong Vocational interest Blank. Have further conference after these results are in. Might have her try Work-limit intelligence test.

All college ave. has been: 1.16-1.00-1.07-1.22-1.34

Chart made and reported by E. B. V. O.

Date 3-10-39

Figure 16. A Profile of a College Student for Use in Guidance.

profile in use at the Pennsylvania State College.[10] Both the raw scores and the percentile values are given, as well as the nature of the population upon which they were based. It will be noted, for example, that on group intelligence tests this young woman falls below the 10th percentile in the select group of which she is a member, but is average or better when compared with high-school seniors and the general adult population.

Her score on the Inglis vocabulary test suggests that one of her chief difficulties is that she has a very limited vocabulary in the fields of the general intellectual reader. There is also evidence from her performance on several of the tests that there is another deficiency in manipulating verbal relations. It is recommended that she go to the Reading Clinic for further study of her reading difficulties. The recommendation that she take the Strong Vocational Interest Blank is to determine her likes and dislikes for clerical work as compared with teaching. It will be noted that her scores are very low on the social and emotional adjustment sections of the Bell Adjustment Inventory, but are much better on the clerical and mechanical aptitude tests. It is thought possible that when all the evidence is available she may wish to modify her present vocational plan to become a teacher of Spanish and English.

The Lake View High School in Chicago prepares a somewhat similar profile for each graduating senior.[11] By using these profiles the homeroom teacher is able to have the pupils evaluate themselves objectively. The originals are filed in the guidance folders, but interested pupils often make copies for themselves.

In constructing and interpreting profiles one caution must be observed. *All tests used in the profiles have to be standardized on similar groups.* If this is not done, as will be apparent from the discussion in the next chapter, the peaks and valleys in the curve may reflect differences in the norms on the tests rather than the strengths and weaknesses of the pupil. The safest profiles are those based on test batteries or the various sections of reasonably long tests.

C. Representing a Frequency Distribution

The ordinary frequency table or distribution does not give a very clear picture of the situation. There are three common methods of representing graphically a distribution of scores: the *histogram* or *column diagram,* the *frequency polygon,* and the *smooth curve.*

---

[10] Edward B. Van Omer and Clarence O. Williams, *Elementary Statistics for Students of Education and Psychology,* page 48. New York: Longmans, Green & Co., 1940.

[11] Clifford E. Erickson and Marian Crossley Happ, *Guidance Practices at Work,* pages 235–237. New York: McGraw-Hill Book Company, Inc., 1946.

The histogram or column diagram. The *histogram* is a series of columns, each of which has as its base one class interval and as its height the number of cases, or frequency, in that class. Figure 17 represents a histogram showing the distribution of percentage values assigned to an arithmetic paper by forty-two scorers. As the greatest frequency is 9, at the 60– class, it is not necessary to extend the vertical or frequency scale at the left beyond 9. As the scores range from the 30– class to the 75– class, it is necessary to represent the horizontal scale only through that distance. It is customary, however, to extend the scale one class interval above and below that range, and to indicate the omission of part of the scale between 0 and the lowest class by a broken line. In order to avoid having the figure too flat or too steep, it is usually well to arrange the scales so that the width of the figure is about one and one half times its height. In actual practice it is customary to represent the histogram in outline form, rather than to show the full length of the columns. Figure 18 illustrates the outline form of the histogram. Figure 19 is an interesting use of the histogram in which each pupil is represented by a code number in an appropriate square. The data are the mental ages of the pupils, whose test records appear on page 266.

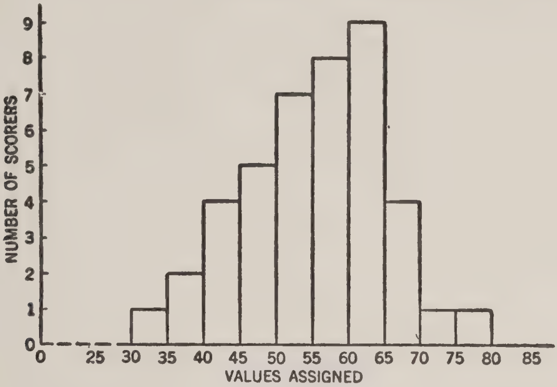


Figure 17. A Histogram, or Column Diagram, Representing the Percentage Values Assigned to an Arithmetic Paper by Forty-Two Scorers.

The frequency polygon. The process of constructing the *frequency polygon,* usually termed merely the *polygon,* is very much like that of constructing the histogram. In the histogram the top of each column is indicated by a horizontal line the length of one class interval, placed at the proper height to represent the frequency at that class. But in the polygon a point is located above the mid-point of each class interval and at the proper height to represent the frequency at that class. These points are then joined by straight lines. As the frequency is zero at the classes above and below those in the distribution, the polygon is completed by connecting the points that represent the highest and lowest classes, with the base line at the mid-points of the class intervals next above and

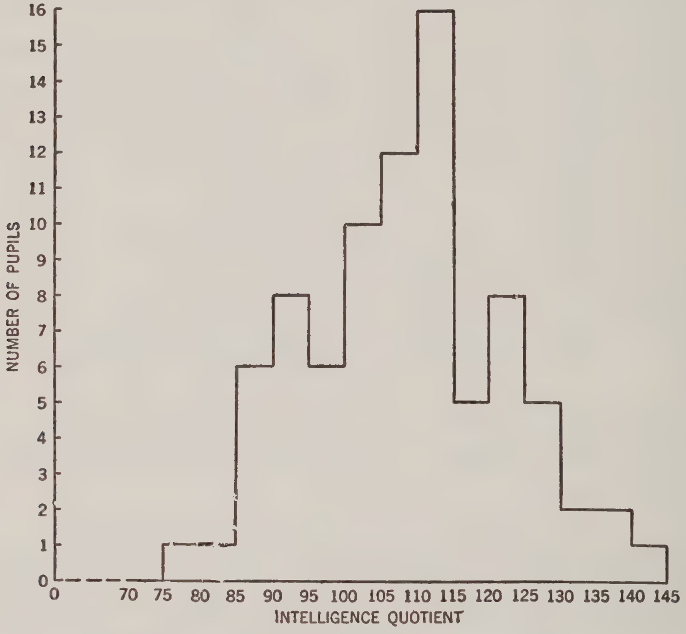


Figure 18. A Histogram, or Column Diagram, Representing the Distribution of IQ's in a Small Junior High School.

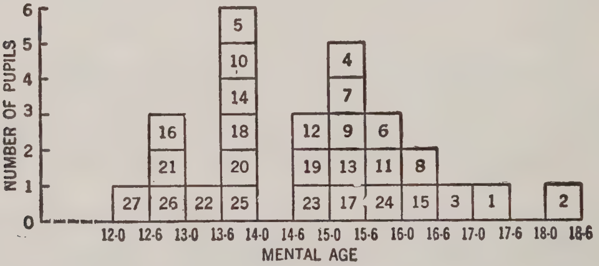


Figure 19. The Distribution of Mental Ages in an Eighth-Grade Class of Twenty-Seven Pupils.

below. Figure 20 shows a polygon for the same data represented by a histogram in Figure 17.

The smooth curve. Sometimes a *smooth curve* is drawn instead of the frequency polygon. The points are located in the same manner for both. The only difference is that for the former a smooth curve is drawn through the points, and for the latter a broken line is used. The most common use in educational measurement of the smooth curve is in the so-called *normal curve,* or *probability curve.* Figure 21 shows such a curve superimposed upon a

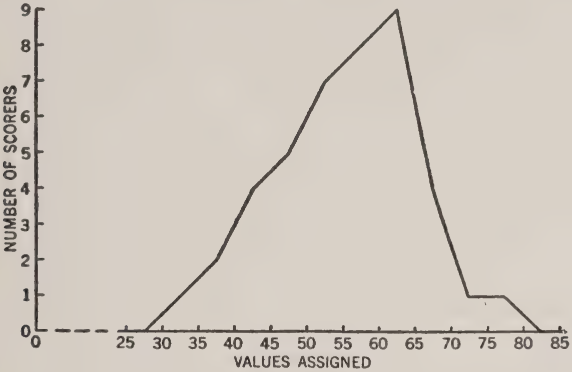


Figure 20. A Frequency Polygon Representing the Percentage Values Assigned to an Arithmetic Paper by Forty-Two Scorers.

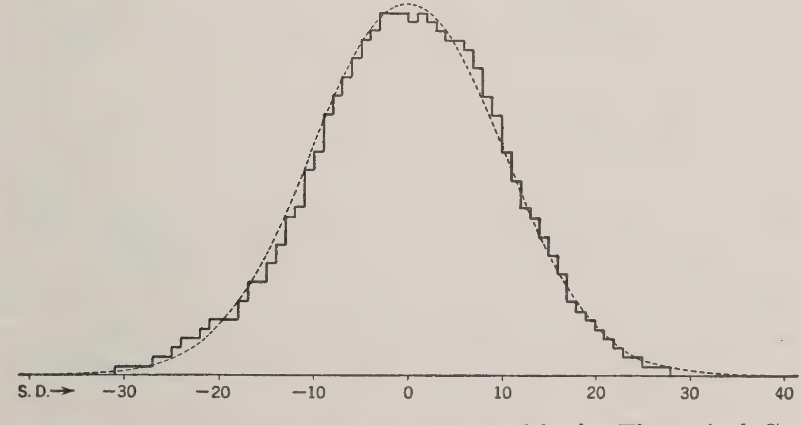
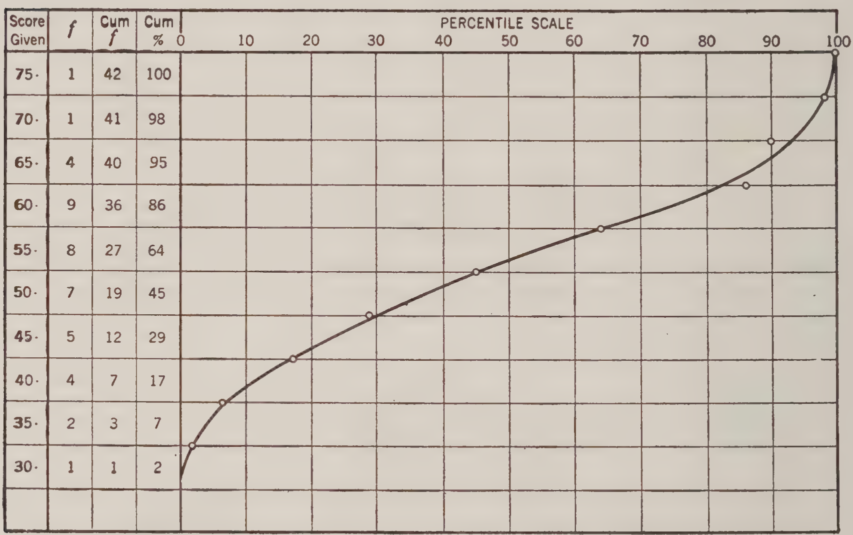


Figure 21. An Actual Curve Compared with the Theoretical Curve of Probability. Actual curve is based upon single curves for eleven well-known group intelligence tests administered to the ninth grade. (From Thorndike's *The Measurement of Intelligence,* Bureau of Publications, Teachers College, Columbia University, page 529.)

histogram representing the actual distribution of ninth-grade pupils on eleven intelligence tests.

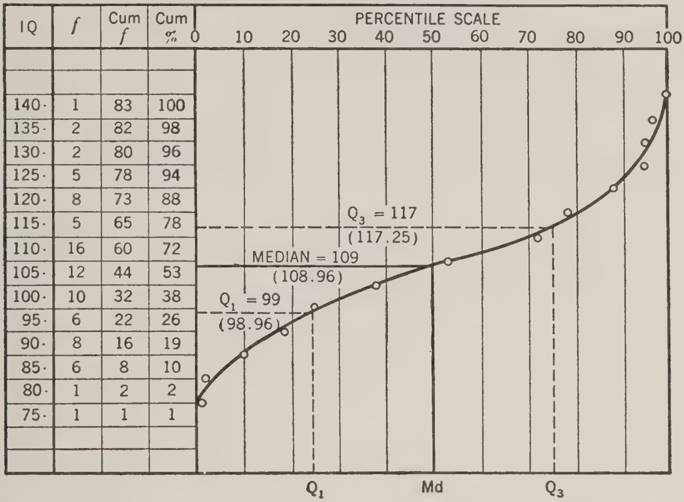
There is one smooth curve, however, which is widely used in representing test scores. This is the *percentile curve,* or *ogive.* Figure 22 shows the percentile curve used to represent the data already employed to illustrate the histogram and the polygon. It will be noted that the scales in the percentile curve are in the reverse order to their position in the histogram and polygon; that is, the score is indicated on the vertical axis and the frequency is indicated on the horizontal axis. The points that determine the percentile curve are located on the horizontal line at the upper limit of each class, at the position that indicates on the horizontal scale the percentage of scores up to and including that class. It will be noted, also, that two columns have been added to the ordinary frequency table. The cumulative frequency column indicates the number of scores up to and including each class. For example, there is one score in the 30– class, and there are two in the 35– class, making a cumulative frequency of 3 in the two lowest classes. The cumulative per cent column shows what per cent each of these cumulative frequencies is of the total. In the illustration the total, N, is 42. The first entry in this column is, of course, 100; the second is 98, because 41 is 98 per cent of 42; the third is 95, because 40 is 95 per cent of 42;



Score Given	f	Cum f	Cum %
75-	1	42	100
70-	1	41	98
65-	4	40	95
60-	9	36	86
55-	8	27	64
50-	7	19	45
45-	5	12	29
40-	4	7	17
35-	2	3	7
30-	1	1	2

Figure 22. A Percentile Curve Representing the Percentage Values Assigned to an Arithmetic Paper by Forty-Two Scorers.

and so on for the others. Each value in the cumulative per cent column is represented as a point on the upper limit of that class interval, since it includes the percentage of scores up through that class. These points determine the curve. As a rule, especially in small groups where irregularities are most likely to occur, it is best to miss some of the points in order to obtain a smooth and regular curve; but care should be exercised in order to leave about as many points on one side of the line as on the other. Figure 23 shows a curve that does not touch all the points. Otis[12] suggests that such a smooth curve, although it does not exactly represent the actual sampling, probably indicates very closely what is to be expected "in the long run."



IQ	f	Cum f	Cum %
140·	1	83	100
135·	2	82	98
130·	2	80	96
125·	5	78	94
120·	8	73	88
115·	5	65	78
110·	16	60	72
105·	12	44	53
100·	10	32	38
95·	6	22	26
90·	8	16	19
85·	6	8	10
80·	1	2	2
75·	1	1	1

Figure 23. A Percentile Curve Representing the Distribution of IQ's in a Small Junior High School. The Values of Q_1, Median, and Q_3 Read from the Curve Are Shown with the Computed Values (in parentheses).

Normal and skewed curves. Regardless of whether a distribution is represented as a histogram, a polygon, or a smooth curve, the curve will be either symmetrical in shape, or else pushed or pulled to the right or left. A symmetrical curve that is balanced in the center and slopes regularly in both directions is said to be *normal.* One that is pushed or pulled in one direction is said to be *skewed.* If the peak of the curve is toward the upper end of the scale, with the longest slope downward toward the lower end of the scale, the curve is negatively skewed. On the other hand, if the peak of the curve is toward the lower end of the scale, with the longest slope toward the higher end of the scale, the curve is positively skewed. Both kinds of curves are shown in Figure 24. Many curves met with in educational measurement show skewness, although the departure from the normal bell-shaped curve is not very great in larger samplings unless some selective factors are operating.

---

[12] Arthur S. Otis, *Statistical Method in Educational Measurement,* pages 79–80. Yonkers: World Book Company, 1925.

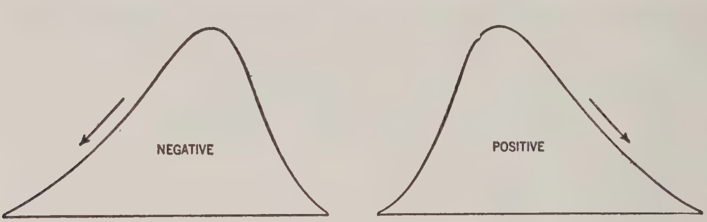


Figure 24. Negative and Positive Skewness.

Other graphs. A *bar graph* is often useful for representing frequency distributions and other data. The principal difference between a bar graph and a histogram is that the bars in the former are not so wide as the class intervals, while the columns in the latter are the full width of the class intervals. The bar graph is also frequently drawn horizontally rather than vertically. Figures 11 and 12 are examples. Satisfactory bar graphs can often be made on the typewriter. Figure 25 and Figure 26 illustrate two such bar graphs. Other graphs, such as the *circle,* or *pie graph,* and various *picture graphs,* or *pictographs,* are occasionally met with in educational measurement.

Which graph is best? As is to be expected, no one type of graph is equally good for all purposes. The histogram is the easiest of all to understand and is usually best if but one distribution is being

```
 IQ       f    |
               |
145-      1    | X
140-      2    | XX
135-      2    | XX
130-      5    | XXXXX
125-      8    | XXXXXXXX
120-      5    | XXXXX
115-     16    | XXXXXXXXXXXXXXXX
110-     12    | XXXXXXXXXXXX
105-     10    | XXXXXXXXXX
100-      8    | XXXXXXXX
 95-      6    | XXXXXX
 90-      8    | XXXXXXXX
 85-      6    | XXXXXX
 80-      1    | X
 75-      1    | X
```

Figure 25. Bar Graph Made on the Typewriter, Showing the Distribution of IQ's in a Small Junior High School.

represented. If two or more distributions are to be compared, however, polygons are usually better, since so many lines coincide when histograms are superimposed that the picture is likely to be confusing. To avoid this difficulty a series of histograms are sometimes placed one above the other. Figure 50 on page 460 is a good example. The percentile curve has many advantages not possessed by other curves. The first of these is that it is possible to estimate with a high degree of accuracy the quartiles, medians, and other similar points. This means that one can read directly from the curve both the central tendency and the variability of the distribution. This is illustrated in Figure 23. As will be shown in the next section by means of percentile curves, several groups can be presented, for convenient comparison, on a single sheet. The principal value of bar graphs, circle graphs, and picture graphs lies probably in school publicity and in the motivation of learning. "A successful graph," as Scates points out, "depends far more on careful thought and judgment than on techniques." [13]

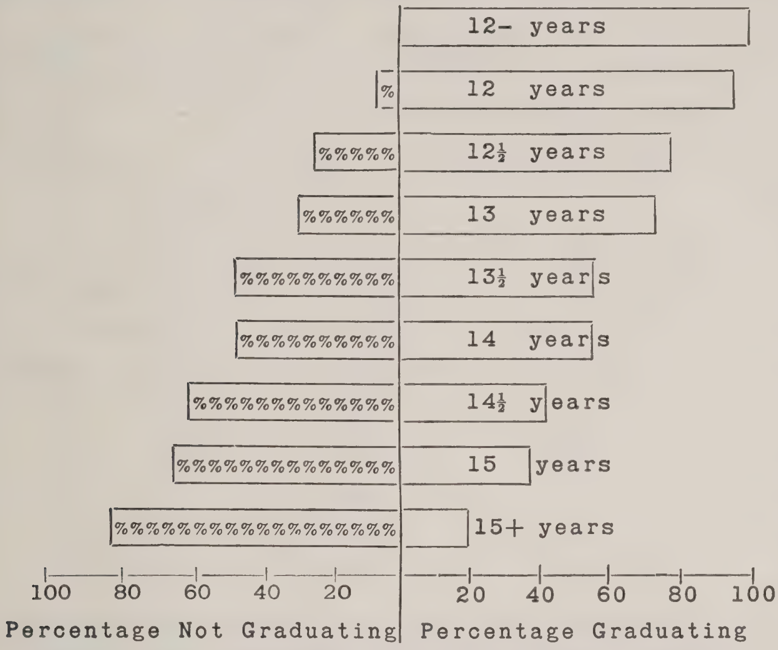


Figure 26. Bar Graph Made on the Typewriter, Showing the Percentage of Pupils of Each Age Group Who Graduate from High School and the Percentage Who Enter High School but Do Not Graduate.

---

[13] Douglas E. Scates, *op. cit.*, page 568.

D. Representing Two or More Distributions

There are many occasions when it is desirable to compare two or more distributions. School administrators may wish to compare the intelligence or achievement of the pupils in various classrooms or buildings. The overlapping among the various grades within a single building is a striking way to present the need for reclassification and sectioning.

Representing entire distributions. When it is important to compare two or more entire distributions, as would be the case in a study of the classification status of a school or school system, the choice will usually lie between the frequency polygon and the percentile curve. The difficulty of superimposing two or more histograms has already been pointed out. A series of polygons may be drawn on the same sheet one above the other, or alongside each other. Figure 27 illustrates a method of showing overlapping by bar graphs made on the typewriter.

Score	Grade					
	7	8	9	Seventh	Eighth	Ninth
200-			3			999
180-	1	4	5	7	8888	99999
160-	3	3	7	777	888	9999999
140-	4	9	7	7777	888888888	9999999
120-	11	7	11	77777777777	8888888	99999999999
100-	4	7	2	7777	8888888	99
80-	4	2	1	7777	88	9
60-	1	3		7	888	
40-		1			8	
20-		1			8	

Figure 27. Graph Made on the Typewriter, Showing the Overlapping of Grades Seven, Eight, and Nine in Reading Comprehension.

The use of polygons. The distinct advantage of polygons over histograms and bar graphs for representing a series of distributions is that polygons can be superimposed upon each other. In this form comparisons among distributions are more easily made. Figure 28 illustrates this possibility with the distribution of reading comprehension scores on the Iowa Silent Reading Test for the seventh, eighth, and ninth grades of a certain school. One fact stands out clearly, the great overlapping of the three grades in reading ability. But even with only three distributions the lines cross and recross so many times as to make any accurate comparison of one grade with another somewhat difficult. More than three classes can hardly be represented in the same graph by frequency polygons without considerable confusion. It is also difficult to compare distributions where the numbers of cases vary greatly, unless each frequency is represented as a per cent of its total.

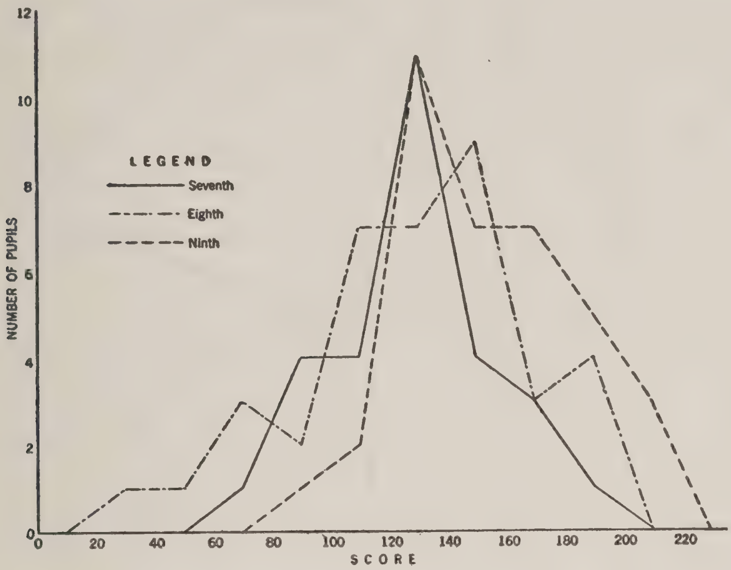


Figure 28. Frequency Polygons Representing the Distribution in Reading Comprehension on the Iowa Silent Reading Tests for the Seventh, Eighth, and Ninth Grades of a Certain School.

The use of percentile curves. For the graphic comparison of two or more distributions the percentile curve has certain outstanding advantages. Since the frequencies are reduced to per cents, it is readily possible to compare groups of unequal size. Another important advantage is that several distributions can be represented in a single graph without difficulty or confusion. Figure 29 shows the distribution of reading comprehension scores for the same grades as in Figure 28 in the form of a percentile curve.

From these percentile curves several relationships are observable that were not apparent in the polygons. It is quite clear that although the seventh and eighth grades have almost exactly the same average scores, the eighth grade has greater variability. This is evident from the fact that the upper half of the eighth grade exceeds the upper half of the seventh grade, but that the lower half of the eighth grade falls behind the lower half of the seventh.

Furthermore, although the ninth grade runs rather consistently above the other two grades, about 25 per cent of the ninth-grade pupils fall below the median of the seventh and eighth grades.

Representing central tendencies of a series of distributions. It is frequently necessary to represent, not the entire distribution, but only the central tendencies or averages. A learning or progress curve is an illustration. Figure 30 shows a graphic picture of the results of a learning experiment. It shows three groups, one with no knowledge of progress, one with partial knowledge of progress, and one with full knowledge of progress. It will be noted that after the second trial the progress was roughly proportional to the amount of knowledge possessed. A simple line graph makes this clear.

Test	Seventh			Eighth			Ninth		
Score	f	Cum f	Cum %	f	Cum f	Cum %	f	Cum f	Cum %
200							3	36	100
180	1	28	100	4	37	100	5	33	92
160	3	27	96	3	33	89	7	28	78
140	4	24	86	9	30	81	7	21	59
120	11	20	71	7	21	57	11	14	39
100	4	9	32	7	14	38	2	3	8
80	4	5	18	2	7	19	1	1	3
60	1	1	4	3	5	14			
40				1	2	5			
20				1	1	3			

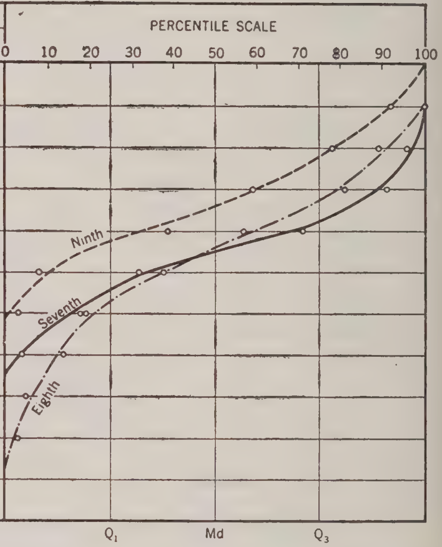


Figure 29. Total Comprehension Scores on the Iowa Silent Reading Tests for the Seventh, Eighth, and Ninth Grades.

Another common use of the line graph is for comparing two or more schools through several grades, or of one school with the norms on a test. Figure 31 shows the correct and the incorrect construction of such a graph. The solid line connects the median scores on a reading test for grades four to nine, inclusive. The tests were given in October, or one tenth of the way through the grade. The dash line connects the norms *incorrectly drawn as of the date when* *the tests were given.* But the norms in the manual are for the *end of the grade.* The dot-dash line connects the norms at the proper grade location. It will be noted that when the line is incorrectly

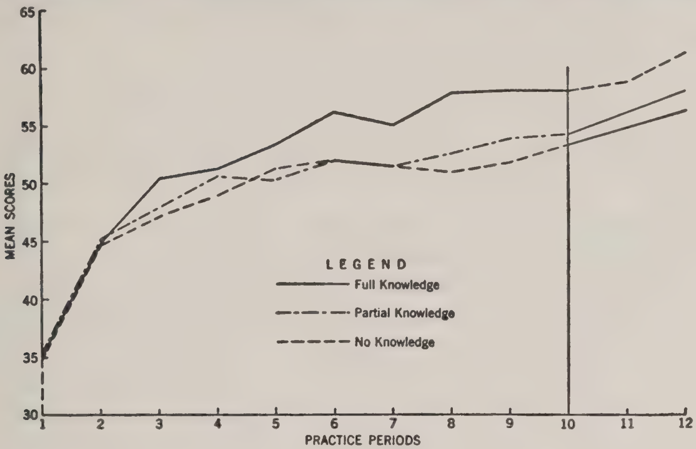


Figure 30. The Learning of Three Groups Compared, One with Full Knowledge of Progress, One with Partial Knowledge of Progress, and One with No Knowledge of Progress.

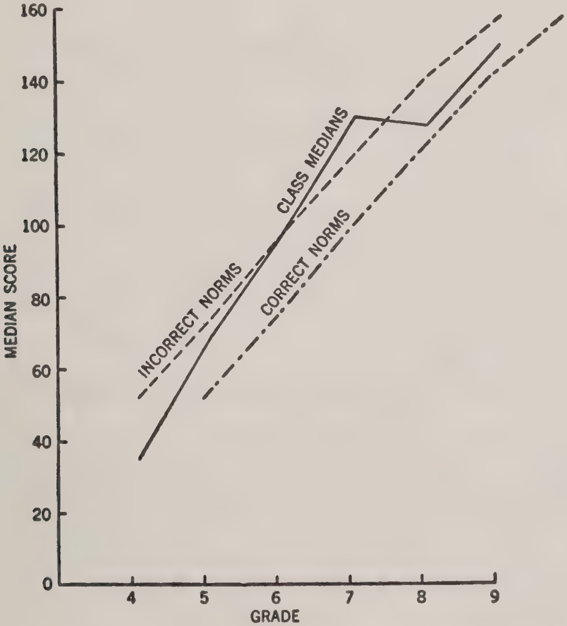


Figure 31. Correct and Incorrect Location of the Norms in a Line Chart, Showing Median Scores on a Reading Test.

located only the seventh grade appears to exceed the norm, whereas in reality every grade does. The horizontal axis should be considered a scale and the points determining the lines should be located carefully with reference to it. This principle is often disregarded. The data upon which this graph is based are:

	GRADES					
	4th	*5th*	*6th*	*7th*	*8th*	*9th*
School Medians for October	35	68	98	130	128	150
Norms for End of the Year	53	74	98	120	141	158

The figure presents much more clearly than do the numerical data the comparison of the school with the norms.

Figure 32 shows the profiles for the third, fourth, fifth and sixth grades of a certain school made by connecting the median scores on each part of the New Stanford Achievement Tests. This figure shows clearly that the school is weak in spelling (dictation) and arithmetic computation, and particularly strong in literature and the social studies. It is evident that this school is stressing the content subjects at the expense of some of the more formal tool subjects. Whether or not this appears to be a desirable emphasis depends upon one's philosophy of education. There is also evidence that the increments of progress are becoming less and less as the pupils advance through the grades.

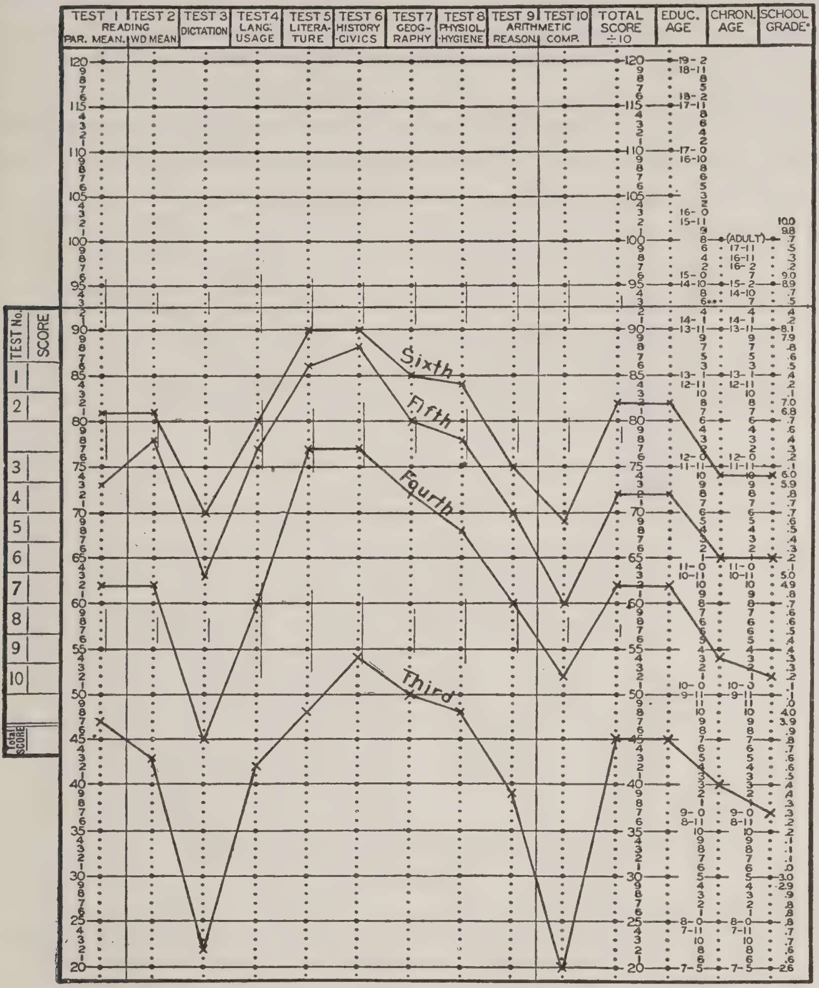
Representing the central tendencies and variabilities of a series of distributions. The variabilities, as well as the central tendencies, of a series of distributions may be shown in a similar manner by line graphs. Figure 33 is an illustration. This figure shows the median, Q_1, and Q_3, for each grade from four to nine, inclusive, in reading comprehension. While the three lines have the same general shape, they converge slightly at the seventh grade, where the variability is least. It would be possible to include from the table of norms the corresponding medians and quartiles for the typical school, but to do so would make the figure too complicated for easy interpretation.

Figure 34 is a bar graph which shows the central tendency and variability in educational age of grades 2B to 9A, inclusive, in a small city school system.[14] In each grade the vertical line indicates the total range, the vertical bar indicates the range of the middle 50 per cent, and the middle of the bar is the approximate position of the median. The horizontal lines across the full width of the graph indicate the norms for the beginning of each grade. It will be noted

---

[14] *Report of the Public Schools of Shelbyville, Kentucky,* page 73. Bulletin of the Bureau of School Service, Vol. I, No. 1. Lexington: University of Kentucky, 1928.

EDUCATIONAL PROFILE CHART: NEW STANFORD ACHIEVEMENT TEST, ADVANCED EXAMINATION



* Grade defined as in Table 1 of the *Directions for Administering*. ** Educational Ages above this point are extrapolated values. See *Guide for Interpreting* for explanation of vertical lines.

☞ *This Profile Chart is the table of norms for the Advanced Examination.*

Figure 32. Grade Profiles for the Third, Fourth, Fifth, and Sixth Grades of a Certain School Made by Connecting the Median Scores on Each Part of the New Stanford Achievement Tests. (Published by World Book Company, 1929.)

that the part of each bar that is crosshatched indicates the proportion that is overage, or above the norm, while the shaded part is the proportion that is underage, or below the norm. The overlapping is especially marked from 7B to 9B. In fact, very little progress is shown from 7B to 9B. This condition suggests the advisability of trying to find out whether these ninth-grade classes happened to be weaker than usual, or whether the teaching emphasis is responsible for the apparent lack of improvement. This type of graph is an effective means of presenting the essential features of a total situation. Here the amount of overlapping is most impressive. It will be noted, for example, that those whose EA is 12–6 are found in all grades from 5B to 9A, and that pupils classified in 8A vary in EA from the 4A level to the 10B level.

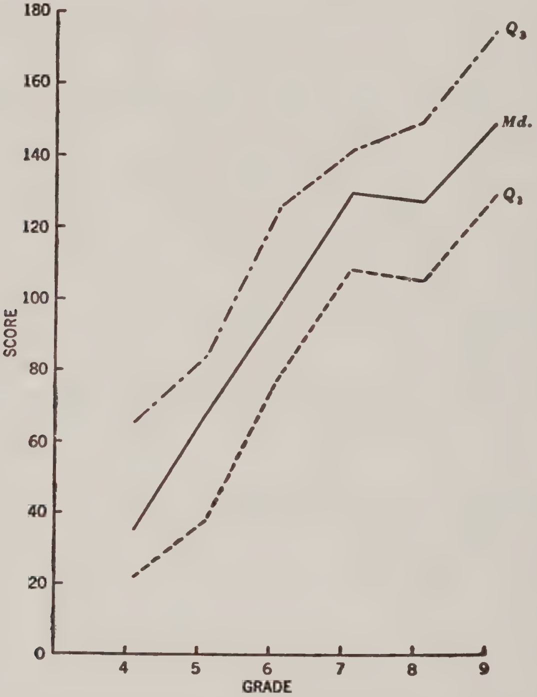


Figure 33. A Line Graph Showing the Medians and Quartiles for Grades Four to Nine, Inclusive, in Reading Comprehension.

E. General Suggestions for Constructing Graphs

Varied practice. A wide diversity of practice will be found in the construction of graphs as used in psychology and education. The title is sometimes placed above the graph, but it is better practice to place it below. In nearly all books and periodicals the graph title is placed below, but in unpublished charts such as wall charts the title is often more effective when lettered above; see Figure 51 on page 466. The figures are numbered consecutively with Arabic numerals placed at the beginning of the title. Sometimes the title is written in capital letters, as in tables; sometimes the initial letters of all important words are capitals; and again, only the first word

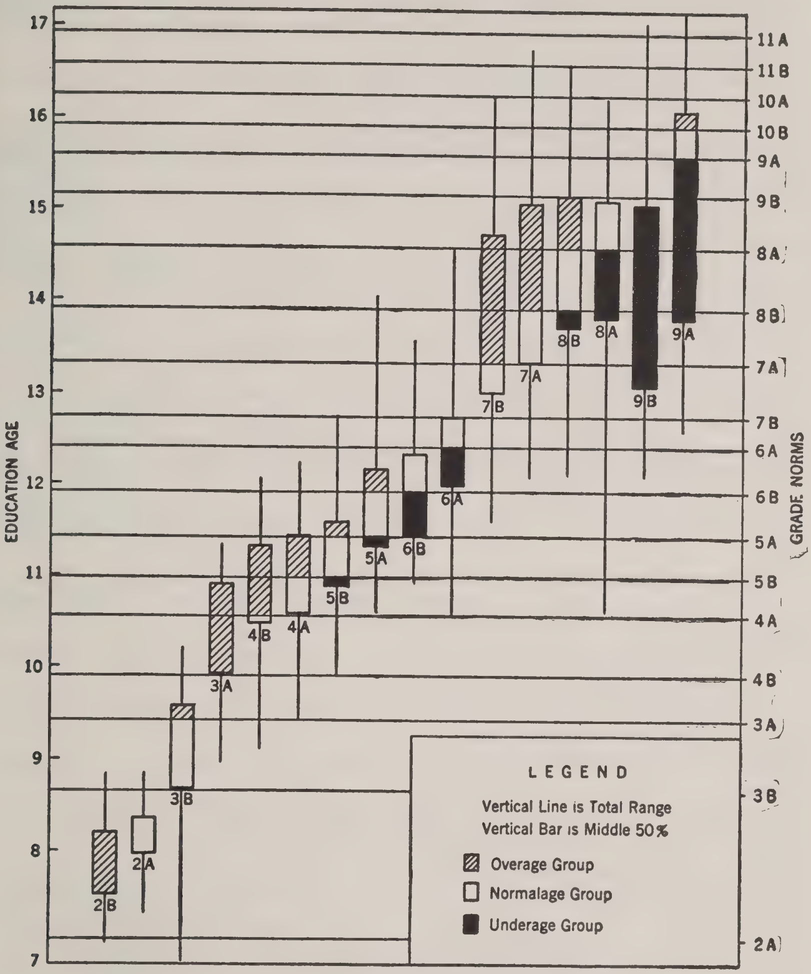


Figure 34. The Central Tendency and Variability in Educational Age of Grades 2B to 9A, Inclusive, in a Small City School System.

in the title is capitalized, unless there are proper names, in which case the usual rules for capitalization apply. The second of these methods is perhaps most common.

Suggested standards. Several years ago a committee composed of representatives of the various groups interested in graphical methods prepared a report[15] recommending certain standards for constructing graphs. This report covers most of the points required for the proper representation of educational data. The following rules are taken from the report:

1. The general arrangement of a diagram should proceed from left to right.
2. Where possible represent quantities by linear magnitudes, as areas or volumes are more likely to be misinterpreted.
3. For a curve, the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.
4. If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.
5. The zero lines of the scales for a curve should be sharply distinguished from the other coördinate lines.
6. For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100 per cent line or other line used as a basis of comparison.
7. When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning or end of time.
8. When curves are drawn on logarithmic coördinates, the limiting lines of the diagram should each be at some power of ten on the logarithmic scales.
9. It is advisable not to show any more coördinate lines than necessary to guide the eye in reading the diagram.
10. The curve lines of a diagram should be sharply distinguished from the ruling.
11. In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the curves representing the separate observations.
12. The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.
13. Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.
14. It is often desirable to include in the diagram the numerical data or formulae represented.
15. If numerical data are not included in the diagram, it is desirable to give the data in tabular form accompanying the diagram.
16. All lettering and all figures on a diagram should be placed so as to be easily read from the base as the bottom, or from the right-hand edge of the diagram as the bottom.
17. The title of a diagram should be made as clear and complete as possible. Sub-titles or descriptions should be added if necessary to insure clearness.

A useful manual which treats of the different phases of the construction of line charts has been prepared by the Committee on Standards for Graphic Presentation.[16] For a fuller discussion of the general problem of graphical representation, several excellent books are available. Of these the most complete treatment from the standpoint of education is that of Williams, and the most complete general treatment is that of Karsten. For a detailed discussion of the construction and use of percentile curves, the reader is referred to Otis.

---

[15] W. C. Brinton, Chairman, "Preliminary Report, Joint Committee on Standards of Graphic Representation," *Quarterly Publications of the American Statistical Association,* 14: 790–797, 1915.

[16] *Time Series Charts: A Manual of Design and Construction,* 68 pages. New York: American Society of Mechanical Engineers, 1938.

SELECTED REFERENCES FOR FURTHER READING

Alexander, Carter, *School Statistics and Publicity*. Boston: Silver, Burdett and Company, 1919. Chapter XI.

Arkin, Hubert, and Colton, Raymond R., *Graphs: How To Make and Use Them*. New York: Harper & Brothers, 1936. 224 pages.

Brinton, Willard Cope, *Graphic Presentation*. New York: Brinton Associates, 1938. 512 pages.

Karsten, Karl G., *Charts and Graphs*. New York: Prentice-Hall, Inc., 1925. 734 pages.

McCall, William A., *Measurement*. New York: The Macmillan Company, 1939. Chapter XXXI.

Modley, Rudolph, *How to Use Pictorial Statistics*. New York: Harper & Brothers, 1937. 170 pages.

Otis, Arthur S., *Statistical Method in Educational Measurement*. Yonkers: World Book Company, 1925. Chapters V, VII, and IX.

Williams, J. Harold, *Graphic Methods in Education*. Boston: Houghton Mifflin Company, 1924. 319 pages.

CHAPTER X

The Uses and Limitations of Norms

It is self-evident that the value of test scores will be dependent largely upon how well they are understood. The two preceding chapters have considered the summarization of scores by statistical and graphical methods as an aid to their interpretation. The present chapter will consider some closely related problems of interpreting scores by the aid of norms.

A. Norms and Standards

Standardized versus nonstandardized tests. At the outset it is important to make a clear distinction between a *norm* and a *standard.* The terms are frequently used interchangeably, but a distinction should be observed between them. The confusion doubtless arises over the fact that norms are used with standard tests and that a part of the process of standardization is the derivation of norms.

Many standard tests began as informal objective tests made by classroom teachers. When an informal test has gone through the process of standardization, it finally appears as a standard test. It then differs from the original class test in four essential aspects. In the first place, the content has been standardized. This means that each item has survived most careful scrutiny by a competent person, or more likely a group, and that its difficulty and value have been determined by rigid experimental processes that have eliminated its weaker fellows. In the second place, its method of administration has been standardized. This means that definite directions have been worked out, usually with appropriate time limits, and the like. In the third place, the method of scoring has been standardized. This means that scoring keys have been prepared and that definite rules have been formulated for marking the papers and for determining the scores on each part and on the whole test. Finally, the process of interpretation has been standardized, at least in part. This means that tables of norms are now available for interpreting the various scores made on the test. These norms, however, are merely the average scores, usually the medians, which have been made by large numbers of pupils distributed over wide geographical areas and representing various types of schools, and which have been grouped, as a rule, according to chronological age or school grade.

Norms versus standards. The word *standard* implies a *goal* or *objective to be reached.* It should be clear, then, that a *norm* is not a measure of *what ought to be,* a *goal,* but is merely a measure of *what is,* the *status quo.* When a grade or class is up to the norm on the test, it is just an average or typical group.[1] Of course, it may be that this score represents a reasonable performance for the group under the circumstances, but that fact would have to be determined by further inquiry. The mere fact that the grade attains the norm does not of itself establish anything other than that the performance is that of a typical group. Manifestly a group of superior opportunities and capacities ought to make better than a typical record. On the contrary, a group of low ability and opportunity might find it impossible to do that well. Unfortunately, at the present time few tests have more than one set of norms for each grade or age group, all types of pupils and schools being lumped together.[2] What is clearly needed is a norm for at least each major type of school organization and type of pupil. Even then such norms could hardly be regarded as reasonable standards of attainment. For one thing, the norms of achievement tests are never more than tentative. They must be continually changing with increases in length of school term and with improvement in training of teachers, in textbooks, in school equipment, and the like. It is also not unreasonable to assume, human nature being what it is, that the average achievement made with the facilities now available could be considerably better than exists at the present time. In a real sense the only valid norm for the individual pupil is his own past record, and the only valid standard is his maximum capacity for growth.

Reasonable standards, or goals of attainment, are almost altogether lacking. It is conceivable that such standards might be worked out and expressed in numerical units on existing tests, or on others to be devised. But such a process is inherently difficult, whereas the process of building norms is time-consuming and laborious but perfectly simple and straight-forward. In fact, an adequate technique for establishing standards has yet to be worked out. Ideally, a standard would have to be provided for each individual. At any rate, no one standard could be established which would be equally appropriate for everybody, or even for any considerable number. In view of such considerations as these, Wood has said:[3]

---

[1] This statement assumes that the norms are of the usual age or grade type. It is not strictly true for such norms as percentile and sigma scores.

[2] The achievement tests prepared by the Armed Forces Institute have established separate norms for six geographical regions as well as for the country as a whole. See *Educational Record,* 25: 369, October, 1944.

> As currently used, the word *standard* has no place in educational literature outside the perorations of convention orators. . . .
>
> Speaking more constructively, it is sufficient to point out that educational standards are necessarily individual, and in their fundamental nature are akin to the standards of tailors and shoemakers who judge the quality of their products by how well they fit the individual for whom they are intended and who pays for them, and how long they serve him.

A recent writer has satirized the idea of a single uniform standard by imagining what would happen if all the tailors of the country got together and agreed upon a "standard suit."[4] The distressing outcome is described[5] as follows:

> Instead of the old haphazard procedure, the standard suit was brought out when a man went into a tailor shop to get a new suit. If he did not fit the suit, he was rejected then and there. He was thus sentenced to join a nudist colony. Men soon learned that the only thing to do was to eat the right food and take the proper exercise to make them just fit the suit. . . . If he perchance ate something else than that required to make him fit the standard suit, he would be rejected, even though what he ate was better for him from the standpoint of health than that needed to get ready for the standard.

The important thing to remember is that, for the present at least, such standards do not exist in any subject. Certainly an understanding of the way norms are determined would make it obvious that they lay no claims to being goals of performance, unless perchance one is willing to accept mediocrity as a goal.

An attempt has been made by the authors of the Progressive Achievement Tests to provide differentiated standards for different ability levels. In the 1937 revision of the norms they report the median accomplishment of schools whose median IQ's vary from 75 to 114. For example, they report that the average achievement of a class with a median IQ of 110 is about three quarters of a year above the ordinary norms, whereas the average achievement of a class with a median IQ of 85 is the same distance below the ordinary norms. Unfortunately, however, it may be argued on theoretical grounds that such differentiation is more apparent than real.

---

[3] Ben D. Wood, "Basic Considerations," *Review of Educational Research,* 3: 13, February, 1933.

[4] J. N. Swan, "Standardized Tests for Chemistry Teaching," *School and Society* 44: 275–277, August 29, 1936.

[5] *Ibid.,* page 275.

B. Raw Scores and Derived Scores

What a score means. To take a simple case, let us suppose that a certain pupil has made a score of 40 on a spelling test of 50 words. What does this score of 40 mean? To say that the score represents an achievement of 80 per cent is true as far as it goes, but this obvious interpretation leaves much to be desired. As the problem of interpreting a given score in meaningful terms is fundamental in all measurement, it deserves careful consideration.

A *score* on any test is simply a *numerical description of an individual's performance on that test.* A distinction must be made between test performance on the one hand, and ability and capacity on the other hand. Performance is merely evidence of ability or capacity. *Ability* refers to an individual's *actual achievement* at the present time, whereas *capacity* refers to his *potentialities.* Since a test is always a sampling rather than a complete measurement, a pupil's response to the test situation is accepted as an expression of his ability operating under a given set of conditions. But a poor score on a valid achievement test is not necessarily evidence of poor ability in that subject under any and all conditions. It may be due to any number of factors, such as physical illness or discomfort, poor eyesight or hearing, emotional disturbance, or dislike for the teacher or subject.

In like manner, a poor performance on even the best group test of intelligence available is not necessarily positive proof of a lack of what we call "general intelligence." It may be due to any one factor or combination of factors mentioned above as operating in the case of achievement tests. In addition, there are several other factors that may be responsible, such as poor reading ability, inability to understand the English language, and, especially, inadequate learning opportunities in school and outside. For example, Wheeler[6] found that the average intelligence of Tennessee mountain children, as measured by two well-known group tests, is approximately normal at six years, but that it shows a fairly consistent decrease with increases in chronological ages. The data warrant the significant conclusion:

> The general trend of this investigation indicates that the results of both tests are materially affected by environmental factors, and that the mountain children are not as far below the normal as the tests seem to indicate. With the proper environmental changes the mountain children might test near a normal group.

---

[6] L. R. Wheeler, "The Intelligence of East Tennessee Mountain Children," *Journal of Educational Psychology,* 23: 351–370, May, 1932.

Ten years later Wheeler[7] repeated the study in the same region which had shown "definite improvement in the economic, social, and educational status" during the intervening period. Although there was still a tendency for intelligence as measured by the tests to decline in the upper years, the average IQ was ten points higher than it had been a decade earlier.

A study of Kentucky mountain children by Asher[8] reveals similar results, and leads to the conclusion that a valid comparison of the intelligence of urban children and of children in less favorable environments "awaits more adequate measuring methods."

The point is that capacity is always inferred from activity or performance. The inference, for example, that two identical scores on an intelligence test really mean equal degrees of intelligence cannot be safely made unless it is known that the learning opportunities have been at least approximately equal. A full realization of this fact would enjoin more caution than is often shown in the interpretation of scores on so-called tests of general intelligence. Trained examiners exercise care in observing rigidly controlled conditions for administering the tests and objective standards for scoring the papers, but it is often hard to be sure about the pupil's past history, which is reflected to some extent in his present performance.

Raw scores versus derived scores. When a test paper has been marked according to instructions, the score obtained is called a *raw* score or *crude* score. On tests, as distinguished from quality scales, it is often called a *point* score, since the numerical description is in terms of points. On a scale, as for example the Ayres handwriting scale, the numerical description is hardly in terms of points but rather in terms of some arbitrary value assigned to a rank or position. In the example given above, the pupil has a point score of 40 on the spelling test. In other words, 40 describes his performance on that particular test at the time it was administered.

But a raw or point score by itself means very little. It is usually not possible to compare directly a raw score on one test with a raw score on another test, or to add a series of raw scores to obtain a total score. The difficulty is that the units are not comparable. The problem is much like that imposed when adding $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, and $\frac{5}{6}$. It is first necessary to find a common denominator, in this case 12, and then to express all values in terms of that denominator. The problem is then simple:

---

[7] L. R. Wheeler, "A Comparative Study of the Intelligence of East Tennessee Mountain Children," *Journal of Educational Psychology,* 33: 321–334, May, 1942.

[8] E. J. Asher, "The Inadequacy of Current Intelligence Tests for Testing Kentucky Mountain Children," *Journal of Genetic Psychology,* 46: 480–486, June, 1935.

$$\frac{6}{12} + \frac{8}{12} + \frac{9}{12} + \frac{10}{12} = \frac{33}{12}, \text{ or } 2\frac{3}{4}.$$

To meet a similar need, test makers have found it necessary to determine common denominators for their test scores. These are called "derived scores." *A derived score is a numerical description of a pupil's performance in terms of norms.* The norm itself is the performance of a defined group considered to be typical. For example, a pupil who answers correctly 22 questions on the Thorndike-McCall Reading Test has a reading ability which is that of the normal, or average, twelve-year-old child at the end of the fifth grade.

Usually, with standard tests the norms used are either age norms or grade norms. The derived scores merely describe the individual's position in some group. Sometimes the age norms are carried one step further and expressed in terms of quotients; that is, one age score is divided by another. With the exception of quotients, most derived scores are obtained from tables of norms in the test manuals, which give in parallel columns the derived scores equivalent to various point scores. As the problems of interpretation differ somewhat for achievement and intelligence tests, they will be treated separately in the next two sections.

C. The Use of Norms in Interpreting Scores on Intelligence Tests

Mental age versus intelligence quotient. The most commonly used units in which to express the results of an intelligence test are mental age and intelligence quotient, usually abbreviated MA and IQ. It is important to understand the distinction between them. MA is a measure of *mental maturity* and "indicates the level of development which a child has reached at a given time," to use the words of Terman.[9] This degree of mental maturity or level of development is expressed in terms of that *"possessed by the average child of corresponding chronological age."*[10] For example, a point score of 75 on the Terman Group Test of Mental Ability is equivalent to an MA of 13 years and 2 months, usually written 13–2. This means that when the Terman Test had been given to hundreds of children in various parts of the country it was found that the average score of a child with a chronological age (CA) of 13 years and 2 months was 75 points. Any child who makes a score of 75 on this test is said to have an MA of 13–2. But pupils of various CA's make scores of 75 on the Terman Test. It is clear, therefore, that a 10-year-old child with an MA of 13–2 has matured rapidly, whereas a 14-year-old child with an MA of 13–2 has matured at a much slower rate. In other words, MA is a measure of *stage* or *level* of maturity but *not* of *rate.* Rate is measured by the IQ, which is obtained by dividing the MA by the CA.[11] In the preceding illustrations the IQ of the 10-year-old child would be 158 ÷ 120 = 132. It will be noted that the quotient is carried out two places but that the decimal point is omitted. In reality, the IQ is the per cent that the MA is of the CA. In like manner, the IQ for the other pupil is 158 ÷ 168 = 94. The IQ, then, gives us a different interpretation of a score on an intelligence test from that afforded by the MA. The *IQ is a measure of rate of maturity,* whereas the *MA is a measure of level or stage of maturity.* In both cases rate and level are relative to the standardization group. If a child has matured rapidly, he is said to be bright; if he has matured slowly, he is said to be dull. A fuller interpretation is to be given a little later. For the present, it is sufficient to note that ordinarily both the MA and IQ of a pupil should be recorded, for each has its distinctive values —and limitations.

---

[9] Lewis M. Terman, *The Intelligence of School Children,* page 7. Boston: Houghton Mifflin Company, 1919.

[10] *Ibid.*

Advantages of the MA concept. The MA has certain outstanding values. Probably the chief of these is that it makes possible a comparison with achievement scores also expressed in age units, as well as with the CA of the pupil, so long as the derived scores are obtained for the same population or from comparable populations. The age basis of comparison is a much more stable unit than the grade location, which is greatly influenced by the promotion policies of the school. Furthermore, age scores are necessary for determining most of the quotients used in educational measurement.

Limitations of the MA. There are also certain serious limitations of the MA, most of which apply particularly to the use of the concept in the high school. It has often been pointed out that the definition of MA does not hold true for CA's beyond 13 or 14. One reason for this is that the norms were based primarily upon pupils in school, who became an increasingly select group, the weaker ones tending to drop out. This was especially true more than two decades ago when Terman was standardizing the original Stanford-Binet scale, against which practically all later tests have been validated. Then, too, in spite of their appearance, the age units on the scale are probably of unequal length, the annual increments becoming smaller and smaller as they approach maturity, when the curve flattens out altogether. But no way has been devised so far for equating these units, or for making satisfactory allowance for their variation in length. This is the principal reason why true growth curves of mental development are not obtainable up to the present, even when the same individuals have been measured repeatedly over a long period of years. This also complicates the problem of investigating the constancy of the IQ, and of its computation in the later chronological ages. Neither of these limitations, however, is very serious in the elementary school.

---

[11] Usually both MA and CA are expressed in months before dividing.

There is a more serious limitation, which appears to operate on all age levels: namely, that the age units on one test are not fully comparable to those on another test. Miller found,[12] for example, that the mean MA of 57 high-school freshmen whose mean CA was 13–6 varied on ten intelligence tests from 15–8 on the Stanford-Binet to 18–9 on the Miller, Form B. Even on the Miller Form A, the mean MA was 17–7, or less by 1 year and 2 months than on Form B. Of course the variation of individual pupils was very much greater. Baker[13] offers the interesting suggestion that the "discrepancy between group intelligence and Stanford-Binet mental ages is due to the fact that the group tests measure 'area' of intellect and the Binet is more a test of 'altitude' according to Thorndike's description." It is, of course, entirely possible that imperfect standardization is largely responsible. But whatever the explanation, it is clearly necessary in reporting intelligence test scores to indicate both the name of the test and the form used.

It may, of course, be true that under the circumstances the terms MA and IQ are not particularly fortunate when used to describe the scores on existing tests. Be that as it may, the users of these tests should frankly recognize such limitations as exist. It is a curious fact, however, that people are just as loath to recognize the limits of their brain children as are parents to recognize the limits of their flesh-and-blood children. The difficulty of arriving at a rational interpretation of a low score on an intelligence test may as well be due to myopia on the part of the interpreter as to that condition in the parent whose child received the score. Boynton[14] recommends what he calls a "pragmatic attitude" toward the tests, for the facts are that "in a vast majority of cases they work with a high degree of success."

After all, in spite of certain definite limitations, intelligence tests, when intelligently used, do afford valuable information to classroom teachers and school administrators. So long as that is true, whether they measure intelligence or something else, whether the age score is really *mental* age or only *personal* age,[15] would appear to be primarily a matter of academic interest only.

---

[12] W. S. Miller, "The Variation and Significance of Intelligence Quotients Obtained from Group Tests," *Journal of Educational Psychology,* 15: 359–366, September, 1924.

[13] Harry J. Baker, "Intelligence and Its Measurement," *Review of Educational Research,* 5: 195, June, 1935.

[14] Paul L. Boynton, *Intelligence, Its Manifestations and Measurement,* pages 231–234. New York: D. Appleton-Century Company, 1933.

One other limitation of MA and all other gross units is that by lumping together many elements they obscure significant differences. Two children of the same CA might have an MA of 8 years, and yet be quite unlike. One child might be unusually strong in the linguistic elements of the test, but lacking in the more concrete, practical, or common-sense elements, while just the reverse might be true of the other child. This means that the *pattern* of the test responses, as well as the total or average, must be considered. This, of course, does not mean that the total score has no value, but rather that by itself it is inadequate, especially for diagnosis and guidance. The practical suggestion, then, is to consider the pattern as revealed by the profile, as well as the total score, be that an age score or what not. As Thurstone says,[16] "Each individual should be described in terms of a profile of mental abilities instead of by a single index of intelligence."

The computation of the IQ. As ordinarily written, the formula for the IQ is $IQ = \frac{MA}{CA}$. That is, the IQ is the quotient obtained by dividing the mental age of the pupil by his chronological age at the time the test was given. In other words, it is the percentage that the mental age is of the chronological age. As a matter of fact, however, the CA used as a divisor is never more than the age at which the test maker assumes mental maturity is reached. Upon the basis of the evidence available at the time, Terman [17] suggested that a divisor of 16 years be used for all pupils whose CA is 16–0 or above. Experience since that time has been somewhat contradictory. Wells, [18] Pintner,[19] and Boynton [20] conclude that the evidence favors using 14–0, or 168 months, as the maximum divisor. Dearborn [21] offers evidence to support 14–6 as the maximum divisor.

---

[15] *Cf.* Henry C. Morrison, *Basic Principles in Education,* pages 229–232. Boston: Houghton Mifflin Company, 1934.

[16] L. L. Thurstone, "A New Concept of Intelligence and a New Method of Measuring Primary Abilities," *Educational Record,* 17: 133, Supplement No. 10, October, 1936.

[17] Lewis M. Terman, *The Measurement of Intelligence,* pages 140–141. Boston: Houghton Mifflin Company, 1916.

[18] F. L. Wells, *Mental Tests in Clinical Practice,* page 58. Yonkers: World Book Company, 1927.

[19] Rudolph Pintner, *Intelligence Testing, New Edition,* page 83. New York: Henry Holt & Company, 1931.

[20] Paul L. Boynton, *op. cit.,* page 49.

[21] Walter F. Dearborn, *Intelligence Tests,* pages 297–303. Boston: Houghton Mifflin Company, 1928.

In the Revised Stanford-Binet, Terman and Merrill[22] suggest this rule:

> Up to 13–0 the entire C.A. is counted; beyond 16–0, none of it. The C.A. of a subject who is between the ages of 13–0 and 16–0 is counted as 13–0 plus ⅔ of the additional months he has lived. This means that a true C.A. of 14 is counted as 13–8; a true C.A. of 15 as 14–4; and a true C.A. of 16 as 15–0, which is the highest divisor used in the computation of the I.Q.

Table 28 gives the appropriate divisor to use in the computation. This suggestion would appear to be in keeping with the fact that mental maturity is reached gradually rather than abruptly. It is probably true that no test so far using MA and IQ units has been adequately standardized for the upper age groups, and certainly none makes possible a final answer to the vexing problem of the age when mental maturity is reached. Many writers favor using some unit other than IQ, especially beyond the elementary school. The author favors retaining the use of the IQ concept through the senior high school, while making allowances for greater inaccuracies beyond the age of 13.

The actual work of computing the IQ's can be greatly reduced and the accuracy increased by the use of tables such as the Inglis Intelligence Quotient Values[23] or those in Terman and Merrill.[24] A preceding chapter emphasized the need for a careful checking of the scoring and totaling of all scores and the obtaining of the MA or other equivalents from the tables of norms in the manuals. There is one other step in computing the IQ, even with the use of tables, that must be carefully watched to insure accuracy. That is the determining of the CA of the pupil. In the lower grades this age score should be taken from the date of birth as shown on the school records. A young child is likely to put down 9 when he is merely "going on 9," for example. In the intermediate and upper grades it is usually safe to rely on the pupil's answer as given on the test blank. On most tests he is asked to give his age at his last birthday, and then to give the month and day of his birthday, or else to tell how many months it has been since his last birthday. The trouble usually comes with computing the months. This computation should always be checked, preferably by a simple table prepared by the examiner. Table 29 illustrates such a table, prepared for a test given on May 21, from which the months can be read directly. It is desirable to verify the years for those pupils whose

---

[22] Lewis M. Terman and Maud A. Merrill, *Measuring Intelligence,* page 68. Boston: Houghton Mifflin Company, 1937.

[23] Published by World Book Company, Yonkers, New York.

[24] Lewis M. Terman and Maud A. Merrill, *op. cit.,* pages 417–450.

TABLE 28

CA'S TO BE USED IN COMPUTING IQ'S OR IN ESTIMATING MA'S FROM IQ'S COMPUTED AT AN EARLIER DATE

ACTUAL CA	CA TO BE USED (Expressed in Months)
Up to 13–0	Actual CA
13–1 or 13–2	157
13–3	158
13–4 or 13–5	159
13–6	160
13–7 or 13–8	161
13–9	162
13–10 or 13–11	163
14–0	164
14–1 or 14–2	165
14–3	166
14–4 or 14–5	167
14–6	168
14–7 or 14–8	169
14–9	170
14–10 or 14–11	171
15–0	172
15–1 or 15–2	173
15–3	174
15–4 or 15–5	175
15–6	176
15–7 or 15–8	177
15–9	178
15–10 or 15–11	179
16–0 up	180

birthdays come in the month the tests are given or in the next month or so; for even high-school pupils will often make an error of one year. When the correct MA's and CA's are determined, the IQ values can be read from the Inglis or some similar table. If the IQ's are computed by actual division, it is necessary to have the work done twice independently.

Interpretation of the IQ. The IQ is a measure of *brightness,* or of *rate* of intellectual development. Following the lead of Terman, many writers consider IQ's of 90 to 110 as "normal," those below as subnormal, and those above as supernormal. According to this scheme, IQ's below 70 indicate "feeblemindedness." Individuals in this group are often subdivided into three types or levels of feeblemindedness; idiots, below 25; imbeciles, 25 to 49; and morons, 50 to 69, inclusive. Most clinicians recognize these as rather rough and

TABLE 29

A TABLE FOR COMPUTING MONTHS SINCE LAST BIRTHDAY (DATE OF TEST, MAY 21)

BIRTHDAYS BETWEEN DATES	MONTHS SINCE BIRTHDAY
January 6 and February 5	4
February 6 and March 5	3
March 6 and April 5	2
April 6 and May 5	1
May 6 and June 5	0 May 21, Test Date
June 6 and July 5	11
July 6 and August 5	10
August 6 and September 5	9
September 6 and October 5	8
October 6 and November 5	7
November 6 and December 5	6
December 6 and January 5	5

arbitrary groupings and attempt to apply other criteria, such as social sufficiency or success in school. In any case, an individual test would have to be given by a competent person to warrant even a tentative classification of feeblemindedness. Terman has suggested that a minimum IQ of 90 is required for success in the ordinary high school. The percentage of pupils who may be expected to fall above and below certain points in a typical unselected group is given in Figure 35. It will be noted that there are no distinct groups. There is a continuous distribution from the idiot to the genius, and the various degrees of brightness shade into each other until they are as indistinguishable as the colors of the rainbow. It is easy to see that red is different from violet, or to see the difference between red and yellow; but it is hard to tell where orange leaves off and becomes red on the one hand or yellow on the other. The concept of "genius" is worthy of further consideration. Following the lead of Terman, it has been common to interpret any IQ of 140 or above as indicating "genius or near genius." Evidence is accumulating which indicates that this limit is much too low. In an illuminating discussion of this problem Mrs. Hollingworth comes to the conclusion that a minimum IQ of 180 is more defensible, and that works of genius are conditioned by high ability when combined with zeal and the capacity for hard work.[25] Terman supports the conclusion that *"above the IQ level of 140, adult success is largely determined by such factors as social adjustment, emotional stability, and drive to accomplish."* [26]

---

[25] *Thirty-Ninth Yearbook of the National Society for the Study of Education, Part I,* pages 62–63. Bloomington, Illinois: Public School Publishing Company, 1940.

Advantages of the IQ. The identification of various degrees of brightness is one of the advantages of the IQ. Moreover, many studies have shown the IQ to remain relatively constant under ordinary conditions from year to year, although radical changes in the home and school environment, which rarely occur, are likely to be reflected in larger changes in IQ when they do occur. Nemzek [27] summarized 97 studies which used the Stanford-Binet test and 27 studies which used group tests. The median correlation coefficient by the test and retest method was .832 for the individual test and .846 for the group tests. The corresponding range of the middle 50 per cent of the coefficients was .760 to .889 and .779 to .885, respectively. The similarity of results for the individual and group tests is remarkable. But these correlations permit considerable variations, which apparently are more likely to occur at the extremes of the distribution than near the center.[28] There seems to be a tendency for the lower IQ's to decrease somewhat on later tests, while the evidence for the higher IQ's is somewhat contradictory. After six years Terman found that 73 of his "geniuses," still below a CA of 13, had lost in Stanford-Binet IQ, the boys 3 points and the girls 13 points on the average. On the other hand, Cattell at Harvard found that children with IQ's above 120 gained approximately 8 points in three to six years. Most studies have noted a regressive effect, however. But, of course, most cases tend to cluster rather closely about the center of the distribution, where the IQ is "fairly stable." After a summary of the experimental evidence, Cattell [29] arrives at this practical conclusion:

> The results are reported as evidence of the large changes in the IQ which do occur in ordinary school practice and to emphasize the caution with which the results of a single intelligence test must be interpreted, even though it be an individual examination made by an expert.

Since the IQ of the average pupil is likely to be relatively stable, if originally computed for CA's between 6 and 13 years, his MA at any later age can be estimated with considerable assurance from his present CA and his IQ, no matter when the latter was computed. For example, suppose that a pupil who had an IQ of 95 when in the third grade is now in the fifth grade. His present CA is 10–2, or 122 months. His estimated MA in months is 122 × .95, which is 116 months, or 9–8. Comparisons with achievement test scores expressed in ages can, therefore, be made wherever such comparisons are thought desirable, without the necessity of repeating the intelligence tests at the same time. Although it is desirable to repeat intelligence tests until at least three tests have been given during the pupil's educational career, the tests need not be given at the same time as the achievement tests in order to make comparisons. The IQ from a test given by an expert to pupils in the public school may be regarded as sufficiently constant to make adjustments for a different date fairly safe, at least for a period of two or three years. Table 28 gives the adjusted CA's to be used in estimating IQ's for individuals whose CA's are beyond 13–0.

---

[26] *Ibid.*, page 84. Quoted by permission of the Society.

[27] Claude L. Nemzek, "The Constancy of the I.Q.," *Psychological Bulletin,* 30: 154, February, 1933.

[28] Both statistical evidence and clinical experience seem to agree that the Revised Stanford-Binet is particularly reliable at the lower IQ levels. See Quinn McNemar, *The Revision of the Stanford-Binet Scale,* page 13. Boston: Houghton Mifflin Company, 1942.

[29] Psyche Cattell, "Stanford-Binet IQ Variations," *School and Society,* 45: 615–618, May 1, 1937.

Limitations of the IQ. In common with all units in which test scores are expressed, the IQ suffers from two limitations. The zero point is arbitrary rather than real, and the various units are of unequal length or value. The difference between 60 and 70 is not equal to the difference between 90 and 100, or to the difference between 130 and 140. In the same way it is absurd to say that a pupil whose IQ is 120 is twice as bright as one whose IQ is 60, or half again as bright as one whose IQ is 80. Such interpretations are meaningless. But in this regard the IQ is no worse than are all raw scores and practically all other derived scores. For example, it is obvious that when the thermometer registers 10 degrees below zero it is not twice as cold as when the thermometer registers 5 degrees below.

There is also another serious limitation of the IQ. Many studies have shown that the IQ's on one test are not comparable to those obtained on another test. Table 30, taken from Gates,[30] illustrates this problem in the first grade. Pupil A, for example, would be classified as just average or "normal" on one test and as a "genius" on another. Not only do the scores for the various individual pupils vary, but the average for the class varies greatly. One test would indicate that the class is just an average or normal class, with a mean IQ of 109, whereas another test would indicate that the class is "very superior," with a mean IQ of 129. The remaining seven tests indicate that the class is "superior," with a mean IQ of 113 to 118, depending on the test. That these discrepant results are the rule rather than the exception is corroborated by other studies. Sangren[31] reported similar results for first-grade children in Michigan. Boynton[32] gave 14 different tests to several sixth- and seventh-grade classes within a period of two weeks. Out of a typical class of 24 pupils he found that 6 pupils each varied more than 50 points on these tests. On one test a pupil would be classified as practically a moron, and on another test this same pupil would rate a near genius. In a study by Miller, already discussed, similar variations were found in the high school. It would appear, therefore, that the makers of intelligence tests have not agreed very much better upon the norms for interpreting their tests than have ordinary teachers upon the values for interpreting essay examinations.

---

[30] Arthur I. Gates, "The Unreliability of M.A. and I.Q. Based on Group Tests of General Mental Ability," *Journal of Applied Psychology,* 7: 92–100, March, 1923.

TABLE 30

VARIATIONS IN IQ'S OBTAINED ON SIX TESTS BY PUPILS IN THE FIRST GRADE (AFTER GATES)

PUPIL	INTELLIGENCE TEST						*Mean IQ*	*Mean Dev. from Mean*	*Range*
	Stanford-Binet	*Dearborn Exam. I*	*Otis Primary*	*Kingsbury*	*Haggerty Delta I*	*Myers M.M.*			
A	124	104	135	124	125	144	126	9.0	40
C	120.5	121	144	114	125	131	126	7.7	30
L	120	115	131	127	114	136	124	7.5	22
E	135	115	138	126	108	121	124	9.2	30
O	104	109	139	117	106	165	123	17.0	61
G	133	112	132	116	112	125	122	8.3	21
N	110	111	135	118	125	114	119	7.5	25
I	104	107	140	116	111	135	119	12.5	36
P	121	115	137	98	117.5	117.5	118	7.7	39
D	114	107	122	134	102	122	117	9.2	32
F	110	117	119	111	112	129	116	5.3	19
J	96	99	117	104	107	152	113	14.8	56
M	108	108	125	100	99	123	111	9.2	26
H	92.5	99	119	105	120	115	109	9.5	28
K	103	92	98	98	107	99	100	3.8	15
Highest	135	121	144	134	125	165	126	17.0	61
Lowest	92.5	92	98	98	99	99	100	3.8	15
Mean	113	109	129	114	113	129	118	9.2	32
M.D.	10.1	6.0	7.6	8.9	6.8	12.3	5.4	2.2	9.6

All these studies are agreed that the IQ's from different tests must be equated in order to make them comparable, for even when average IQ's on different tests are close together the extremes are likely to vary widely. One solution which has been proposed is to equate all tests in terms of some widely used test.[33] Kefauver[34] has prepared such a table in which ten tests have been equated in terms of the Terman Group Test of Mental Ability. Table 31, based on Kefauver, shows, for example, that a pupil who gets an IQ of 60 on the Terman Test would have IQ's ranging from 31 to 77 on the other tests. IQ values comparable to Terman IQ's of 75, 100, 125, and 140 are also shown in Table 31. It is worthy of note that the IQ values of different tests by the same author are by no means equivalent. The implication for test users is to record the name of the test and form with the IQ.

---

[31] Paul V. Sangren, "Comparative Validity of Primary Intelligence Tests," *Journal of Applied Psychology,* 13: 394–412, August, 1929.

[32] Paul L. Boynton, *op. cit.,* page 231.

Table 31

IQ EQUIVALENTS TO TERMAN IQ'S ON TEN GROUP INTELLIGENCE TESTS FOR FIVE LEVELS OF ABILITY (after KEFAUVER)

Test	Equivalents to Five Terman IQ's				
	60	*75*	*100*	*125*	*140*
Army Alpha	59	76	104	132	149
Dearborn IIC	69	82	104	125	139
Haggerty Delta 2	45	68	103	140	161
Illinois	61	80	109	139	157
Miller A	31	57	101	145	172
Miller B	56	78	114	151	173
Pressey Classification	46	66	98	129	148
Otis Advanced	77	89	109	128	139
Otis S-A Higher	70	83	103	124	137
Otis S-A Intermediate	66	80	102	123	136

Doubtless the fundamental solution is for all test makers to standardize their tests, whether they aim to measure intelligence or achievement, on a national population so chosen as to conform fully to the mathematical theory of sampling. As long as tests continue to be standardized on samples chosen primarily upon the basis of convenience, even when they involve large numbers and wide geographical areas, there is still no assurance that the samples are truly representative and thus comparable with each other.

Because of its numerous limitations some authorities would abandon the IQ concept altogether. Stoddard,[35] for example, characterizes it as a "myth" pure and simple. No one recognizes the limitations of the IQ more clearly than its friends, as this statement from Terman[36] indicates: "An obtained I.Q. is not only subject to chance errors resulting from inadequate sampling of abilities, but also of numerous other errors, including practice efforts, negativism, or shyness, the personal equation of the examiner, and standardization errors in the test used."

---

[33] W. S. Miller, "Variation of IQ's Obtained from Group Tests," *Journal of Educational Psychology,* 24: 468–474, September, 1933.

[34] Grayson N. Kefauver, "Need of Equating Intelligence Quotients Obtained from 'Group Tests,'" *Journal of Educational Research,* 19: 92–101, February, 1929.

[35] George D. Stoddard, *The Meaning of Intelligence,* page 258. New York: The Macmillan Company, 1943.

All things considered, the author is disposed to agree with Terman and Merrill[37] that the sensible thing to do under the circumstances is to "employ the simplest indices available and as rapidly as possible acquaint teacher, school counselors, social workers, and physicians with their significance and their limitation." The MA and the IQ are examples of such "simple indices." However, amateur test users will do well to remember at all times Hildreth's warning that "no one I.Q. ever indicates exactly any child's tested ability."[38] No matter how obtained, the IQ should never be accepted as the final verdict but rather as a point of departure for further investigation.

Other derived scores. To avoid the difficulties in the MA and IQ, other types of derived scores have been proposed. Of these, the three most common will be considered briefly. It must be apparent at the outset, however, that no norm can be any better than the sample upon which it is based or the measuring instrument employed. Errors of sampling and of measurement cannot be avoided by the simple device of shifting the unit in which to express the norms. The Cooperative tests have moved in this direction by taking as a point of reference the "50-point," the score "intended to represent the score which the average white child in the United States would make at the end of the particular course if he had attended a typical school and had had the usual instruction in the subject in question."[39]

The *Personal Constant* (PC) has been suggested by Heinis of Geneva as a substitute for the IQ. Kuhlmann is so convinced of the merits of this method that he includes a table of Heinis Mental Growth Units, which he recommends in place of the IQ for use with the Kuhlmann-Anderson tests. On the other hand, Cattell[40] found the Heinis PC more constant than the IQ for pupils of low intelligence but not for those of high intelligence. Although a promising approach in its present form, the PC has, at best, apparently solved only half the problem. To date it has not received wide acceptance.

---

[36] *Thirty-Ninth Yearbook of National Society for the Study of Education, op. cit.*, page 466.

[37] Lewis M. Terman and Maud A. Merrill, *op. cit.*, page 29.

[38] Gertrude Hildreth, "Stanford Binet Retests of Gifted Children," *Journal of Educational Research*, 37: 301, December, 1943.

[39] John C. Flanagan, *The Cooperative Achievement Tests, A Bulletin Reporting the Basic Principles and Procedures Used in the Development of Their System of Scaled Scores*, page 19. New York: Cooperative Test Service, 1939.

[40] Psyche Cattell, "The Heinis Personal Constant as a Substitute for the IQ," *Journal of Educational Psychology*, 24: 221–228, March, 1933.

A second substitute, proposed by many writers, is the *percentile score,* sometimes called the *centile score.* A percentile is a description of a pupil's position in a typical age or grade group in terms either of the percentage of pupils who fall below that score, or of the percentage who exceed that score. A percentile score of 50 would, of course, be exactly at the median. In like manner a percentile score of 10 would show that in a typical group only 10 per cent make a poorer score than that, while a percentile score of 90 would mean that only 10 per cent make better scores than that, since 90 per cent fall below. This is a very simple and useful system that is widely used for achievement tests also, but it has two limitations. One limitation is that a percentile score of a given magnitude in one group is not directly comparable with the same percentile score in another group. A 10th-percentile pupil in the freshman class is manifestly not the same as a 10th-percentile pupil in the senior class, for example. A second limitation is that the percentile units are of unequal length. This can perhaps be made clear by an examination of Figure 35, below. It will be noted, for example, that in a typical group an IQ of 70 is a 1st-percentile score and that an IQ of 76, which is an increase of 6 points, is a 3rd-percentile score; but that an IQ change of 6 points, from 85 to 91,

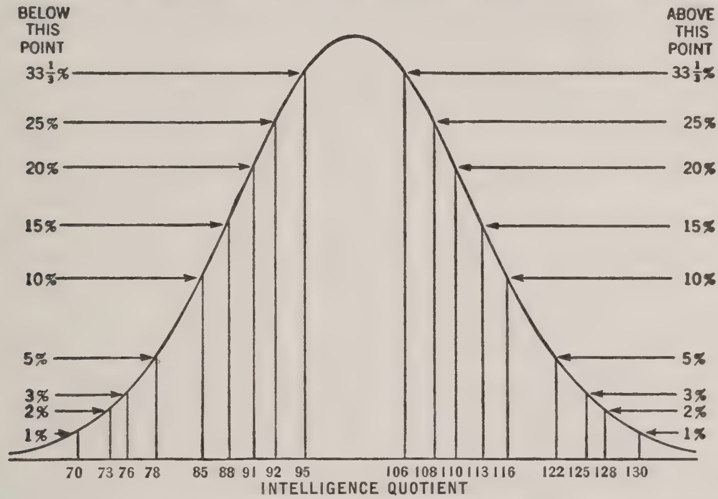


Figure 35. Percentages in a Typical Group Whose IQ's Fall Below and Above Various Points. (Based on Freeman, *Mental Tests,* page 102.)

raises the percentile score 10 points. In other words, the distances between percentiles near the center of the group are much less than those at the extremes.

A third method has been suggested which equates all units on the scale. Several authorities[41] think it is superior to the Heinis PC discussed above. This is the method of *standard scores,* sometimes called sigma scores or z-scores. This method was used by Stutsman in her Merrill-Palmer performance scale. These units are expressed in terms of the mean and standard deviation of the typical age or grade group or, for that matter, of any group.[42] An illustration will help to make the system clear. Suppose that a pupil makes 40 points on one test and 80 points on another test. It is certainly unsafe to say that he did twice as well on one test as on the other, or to combine the two scores into a single score of 120 points. This is evident if we find that the mean score on the first test is 30 points and that the mean score on the second is 90 points. In other words, the pupil is 10 points above average on one test and 10 points below average on the other test. To reduce these scores to a common denominator requires one additional step: namely, to take into consideration the variability of the scores as well as their central tendency. Suppose, then, that the standard deviation of the first is 10 points and that of the second test is 20 points. It is now clear that

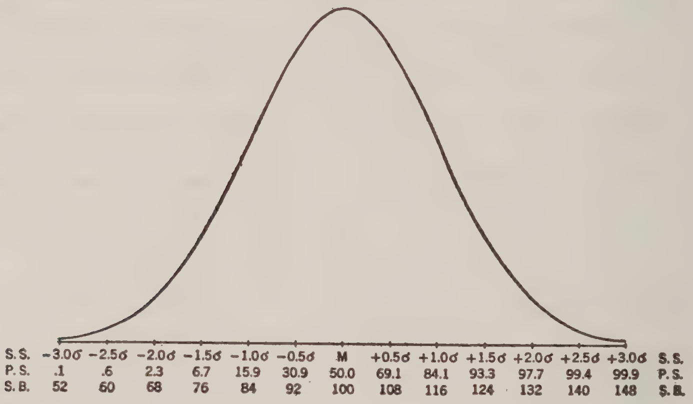


Figure 36. The Relation Between Standard Scores, Percentile Scores, and Revised Stanford-Binet IQ's. (Based on Terman and Merrill, *Measuring Intelligence,* page 42.)

---

[41] Francis N. Maxfield, "Trends in Testing Intelligence," *Educational Research Bulletin,* 15: 134–141, May 13, 1936.

[42] Sometimes the standard score is based upon the quartile deviation instead of the standard deviation. It is then called a C-score.

our pupil is 1.0 standard deviation distance above the mean on one test and .5 standard deviation distance below the mean on the other. These two scores are now standard or sigma scores and are written $+1.0\sigma$ and $-.5\sigma$. To avoid negative numbers, the suggestion is sometimes made that the mean score be called 50 arbitrarily and each standard deviation distance above and below be equivalent to 10 points. In our illustration opposite, the pupil's scores would be 60 and 45.

The system has much to commend it statistically. In fact, its principal limitation is that it appears to be rather cumbersome to handle. That impression is, however, probably due more to its unfamiliarity than to anything else. Some writers[43] point out that not only are MA's and IQ's defined in the usual fashion indeterminate for the upper half of the adult population, but they also argue that standard scores or percentile scores yield much more information even for young children. Figure 36 shows the relation between standard scores, percentile scores, and Revised Stanford-Binet IQ's. It will be noted that the IQ's on the Revised Stanford-Binet may be considered standard scores whose mean is 100 and whose sigma is 16.

Regardless of the type of norms used, the teacher must never lose sight of the fact that all measurement is subject to error and that scores can rarely be taken at face value. Some competent statisticians are so impressed by the "ubiquitous probable error," to use Kelley's phrase, that they think numerical scores of every kind "convey an unwarranted impression of exactitude," and would report the results of intelligence tests in general terms, such as "dull," "normal," or "bright."[44] In the writer's judgment a better practice is to continue to employ the numerical scores but to be aware of their limitations.

D. The Use of Norms in Interpreting Scores on Achievement Tests

Educational age versus educational quotient. In interpreting scores on achievement tests the terms *educational age* and *educational quotient* are widely used in just the same way that *mental age* and *intelligence quotient* are used in interpreting scores on intelligence tests. In other words, educational age, or EA, is a measure of educational maturity, or level or stage of educational growth.

---

[43] L. L. Thurstone and Thelma Guinn Thurstone, "Psychological Examinations, 1940 Norms," *American Council on Education Studies,* 5: 2–3, 1941.

[44] H. E. Garrett, "The Standardization of the Terman-Merrill Revision of the Stanford-Binet Scale," *Psychological Bulletin,* 40: 196, March, 1943. See also *Psychological Bulletin,* 43: 72–76, January, 1946.

In like manner the *educational quotient,* or EQ, is a measure of rate of educational growth or development. The EQ is obtained by dividing the EA by the CA. For example, a 10-year-old boy has made a score of 60 points on a certain achievement test, which is the average score for a 12-year-old pupil. The boy is then said to have an EA of 12–0, which, divided by 10–0, gives him an EQ of 120. In like manner another 10-year-old boy in the same class might make a score of 35 points, which is the average score for a pupil of 8 years and 6 months. His EA is 8–6, and his EQ is 85. It should be noted that the terms EA and EQ refer to scores made on general achievement tests or on test batteries involving several subjects. If a test in only one subject is used, the terms *subject age* and *subject quotient* are employed. For example, a reading test would yield reading ages and reading quotients, while an arithmetic test would yield arithmetic ages and arithmetic quotients, and so on for the other subjects.

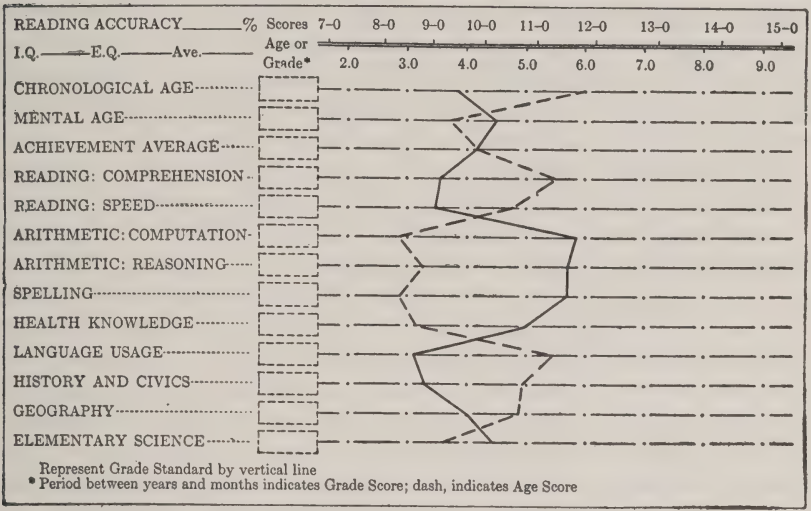
Uses of EA. The value of EA and of the various subject ages is that they make possible a meaningful interpretation of scores in terms of a relatively stable unit, chronological age. They also facilitate important comparisons with norms, on both intelligence and other achievement tests, whenever they have been standardized on comparable groups, as well as with the individual's own MA and CA. The age scores are essential to the determination of quotients of many types.

Limitations of EA. EA and all subject ages have many of the limitations already pointed out in the case of the MA. Probably the most serious is that they reflect the promotion policies and holding power of the schools in which the tests are given. It is a matter of common observation that the performance of a 10-year-old pupil who is retarded in the grade is not the same as that of a 10-year-old pupil who has made normal progress, and much less than that of the accelerated pupil of the same age. Crawford [45] has made an extensive study of the influence of such factors upon norms based on unselected groups, and comes to this significant conclusion: "The factors of chronological age, mental age, and rate of progress affect test norms to a degree that makes the use of norms based on groups in which these are not controlled of doubtful value." His recommendation is that both CA and MA should be used in establishing norms. One solution to the problem is to use only pupils whose MA's are normal for their respective ages in computing norms.

One other limitation of existing age norms is that age units on one test are not comparable to those on other tests that are presumably measuring the same thing. Test publishers owe a service to the public and should prepare tables for equating age norms on various achievement tests in much the same way as has been done by the Cooperative Achievement Tests[46] and the various parts and forms of the Metropolitan and Stanford Achievement Tests.[47] Another much less serious limitation is that the age units on any one scale or test are not necessarily equivalent throughout its length.

---

[45] John R. Crawford, "Age and Progress Factors in Test Norms," *University of Iowa Studies in Education,* 9: 1-39, June 15, 1934.



Represent Grade Standard by vertical line
* Period between years and months indicates Grade Score; dash, indicates Age Score

Figure 37. The Profiles of Two Pupils Who Made the Same Total Score on a General Achievement Test. (From the Modern School Achievement Tests, published by Bureau of Publications, Teachers College, Columbia University.)

An important limitation of the EA and all other gross units is that they lump together many and diverse elements in such a way as often to obscure significant differences. Two pupils of the same CA or MA might have an EA of 10–0, for example. This does not guarantee that they are by any means identical in achievement. One pupil might be greatly accelerated in reading, language, and literature, but retarded in arithmetic, spelling, and science, whereas the exact opposite might be true of the other pupil. EA, which is a composite, or average, has taken no account of the *pattern,* which may afford the key to an adequate interpretation. This fact, of course, does not mean that age scores and other averages have no value, but rather indicates that they are inadequate by themselves. The practical implication is clear. The total score, whether age or what not, is important, but must be considered always in relation to the *pattern* of the responses, usually best represented as a profile. Figure 37, showing two profiles drawn upon the same chart for pupils making the same total score, should make this point clear.

---

[46] Published by Cooperative Test Service.
[47] Published by World Book Company, Yonkers.

Use and limitations of EQ. The method of computing the EQ and the various subject quotients has already been described. As a measure of rate of educational progress these measures are useful in the interpretation of scores on achievement tests. However, no such elaborate scheme for the interpretation of these quotients as was described for the interpretation of quotients on intelligence tests has been worked out. There is nothing which corresponds to such terms as "feeblemindedness" or "genius" on these tests.

EQ's are subject to some, but not to all, of the limitations pointed out for IQ's. Since educational growth continues at least throughout the formal school period, there is no problem of selecting a maximum divisor, such as was described in the case of IQ. It is true that the units are of unequal length and that the quotient technique is not appropriate for use in high school, where age norms on achievement tests are ordinarily not available. Undoubtedly the most serious limitation of quotients, as well as of other norms, is that the units on one test are not directly comparable with those on another test that purports to be measuring the same thing. So far, tables for equating quotients on achievement tests similar to those by Kefauver for intelligence tests have not been published. But they are seriously needed. Better still would be the exercise of greater care in the original standardization. The test record should always indicate the name of the test and the form used, for achievement tests as well as for intelligence tests.

Use and limitations of grade norms. It is a very common practice to interpret achievement tests in terms of *grade norms.* Grade norms on standard tests are usually the average scores made on the test by pupils in each grade when the test has been given to pupils in widely scattered areas. In the earlier tests these grade norms were usually for the end of the grade only, although sometimes for the middle of the grade also. This made comparison with norms somewhat difficult, unless the tests were given at the same time in the year. Figure 31, on page 275, illustrates a simple graphical method of making comparisons with norms for a different time in the year. Of course, such a comparison assumes uniform progress throughout the grade, which may be only approximately true in some instances. A slight variation of such norms for high-school use is to base the norms upon the length of time the subject has been studied rather than upon the grade or year in which it happens to be offered. Since many high-school subjects do not continue over the entire high-school period and have no definite grade location, norms based on the number of semesters the subject has been studied are very useful. The problem of interpretation is just the same as for regular grade norms.

In recent years many tests below the senior high-school level have norms available for every month in the school year. For example, 6.0 means the norm for the beginning of the sixth grade, while 6.5 is the norm for the middle of the grade. In like manner 4.2 means the norm for the fourth grade two months after school starts, and 4.10 means the norm for the end of the fourth grade. Such norms are often called G-scores, and sometimes B-scores. They have the distinct advantage of being readily understood. They also have certain dangers and limitations. For one thing, they tend to imply a degree of mathematical exactness which the accuracy of existing tests hardly warrants. Certainly it is unsafe to take them literally at their face value. A still more serious limitation is the lack of comparability of scores on different tests. Adams found,[48] for example, that eight of the best known arithmetic tests rated the mean performance of 152 pupils all the way from the fifth grade to the eleventh grade, depending upon the test used. It is unnecessary to comment upon the absurdity of fractional-grade norms in a situation like that. The solution is, however, not so much in the abandonment of grade norms as in their further refinement. There is another danger in interpreting grade norms, no matter how accurately determined. This danger arises partly from the fact that a school with an overstrict promotion policy will tend to show up favorably on grade norms simply because of the presence of a great many pupils in the several grades who really belong in higher grades. It is always well, therefore, in any apparently superior school, to make a comparison on the basis of age, to see whether the superiority is real or only illusory. Of course there would be little difference between schools which promote strictly on the basis of CA and those in which the percentages of acceleration and retardation are balanced, a condition which rarely exists.

Ruch and Segel, after noting some evidence that "recent tests may possibly have much more dependable norms than those standardized a decade or so earlier," nevertheless make the suggestion that [49]

---

[48] Summarized by Giles M. Ruch from an unpublished master's thesis by Eunice Adams, *The Comparative Reliability of Eight Arithmetic Tests,* University of California, 1929, in *Review of Educational Research,* 3: 39, February, 1933.

[49] Giles M. Ruch and David Segel, *Minimum Essentials of the Individual Inventory in Guidance,* page 82. Washington: United States Office of Education, 1939.

> . . . many factors peculiar to the individual school system must be considered in the interpretation of tests, such as: the legal age of school entrance, the actual average age of school entrance, rates of acceleration and retardation, rates of elimination from school, percents of failures of pupils, genuine differences in instructional efficiency, and variations in average mental and educational capacity from school to school.

Harris [50] has recently called attention to a rather common error made in the interpretation of grade norms at the primary level. It arises from the failure to take into account the fact that zero performance on an achievement test is 1.0. A first-grade class whose grade score at the end of the year is 2.0 has made only normal progress for the year.

Other norms for achievement tests. Several other types of norms are used, some of which require brief mention. Of these, doubtless the most important are *percentile norms*. As in the case of intelligence tests already discussed, such norms interpret a pupil's score by describing his position in the group in terms of the per cent of pupils who fall below the score made. Generally all percentile values, but sometimes only certain points, such as the 25th, 50th, and 75th, are given. These scores are very easy to interpret, but they have two unique limitations, neither of which is very serious for most purposes. One is that the scale values are unequal in length, and the other is that percentile values in one grade or age group are not directly comparable with those in another. Perhaps percentile norms have their greatest value in interpreting scores on nonstandardized tests.

Standard scores, or *sigma scores,* are also used. They are interpreted in the same manner as are similar scores on intelligence tests. These have already been discussed. McCall has proposed a modification called a *T-score*. The only difference is that the standard group is composed of 12-year-old pupils. All age and grade groups are described by locating their T-score position in this 12-year-old group. The mean is given a value of 50, and each standard deviation distance above and below is divided into tenths, each counting one point. For example, a 15-year-old pupil makes a reading score on the Thorndike-McCall Reading Test, which, according to the table of norms, is a T-score of 60. In other words, this pupil is located 1σ distance above the mean of typical 12-year-old pupils who have taken the test. The T-score technique is sometimes used with other age groups. In that case the ordinary Z-score is multiplied by 10 and then added to 50. The principal limitations of the T-score are that it is not well adapted to high-school tests and that it is rather cumbersome even for grade-school tests. The *C-score* is similar to the T-score, except that the unit is one tenth of a quartile deviation instead of one tenth of a standard deviation. Some publishers employ a sigma percentile graph, or profile.

---

[50] Albert J. Harris, "Note on a Source of Error in Interpreting Grade Norms," *Journal of Educational Research,* 39: 151–153, October, 1945.

Value of local norms. Practically all norms published on tests are so-called *national norms.* When such norms are carefully derived, they are of great value in interpreting the scores. It is easy to overemphasize their value for the ordinary school and school system, however. They must never be taken as standards. There are such wide variations in the length of school terms, in the equipment of schools, in the training and experience of teachers, and in other important respects, among the several states and among the school units of any one state, as to make any single series of norms for the whole nation entirely inadequate. National norms must be supplemented by norms for the state, county, and city school systems, and even for the individual school. What is really important in most cases is the comparison of grades, classes, and schools which operate under approximately the same conditions. Lindquist [51] has recently pointed out several distinct advantages of the regional testing programs used in Iowa for a number of years.

For purposes of classification, what is needed is a set of norms for the school itself. To derive satisfactory local norms, all that is required is to combine all pupils in the same grade and then to compute the median score. If age norms are desired, the pupils will be distributed according to CA or MA, and the medians computed. In like manner percentile scores can be computed from these combined grade and age groups according to the method described on pages 227–233. In larger schools and school systems norms should be derived for slow and rapid learners as well as for average or normal learners on each grade level.

It would appear, then, that the more specific the norm the more useful it becomes. The progressive schools are coming to recognize that each individual personality has its own unique pattern of growth. This position is clearly stated [52] as follows:

> The time has come when we should cease to be primarily interested in comparing one child with another, one class with another, or any class with a norm. We should be primarily interested in comparing each child with himself, with his past record, and with his potentialities. To center attention elsewhere is to miss the point—to miss the service which tests can render.

---

[51] E. F. Lindquist, "Nationally Coordinated Regional Testing Programs," in *New Directions for Measurement and Guidance,* pages 87–103. Washington: American Council on Education, 1944.

[52] Douglas E. Scates, "The Improvement of Classroom Testing," *Review of Educational Research,* 8: 532, January, 1939.

E. Methods of Comparing Intelligence and Achievement

One of the most important questions to raise about any pupil is: How well is he getting along in comparison with his capacity? Whenever intelligence tests and achievement tests for the pupil have been expressed in comparable terms, a rough answer to this question is possible. But as a matter of fact, the problem is far more difficult than would appear on the surface, owing largely to the inadequacies of existing tests. According to Kelley,[53] about 90 per cent of the capacity measured by a so-called general intelligence test, such as the National, is the same as that measured by an all-around achievement test battery, such as the New Stanford, Progressive, or Modern School tests. In like manner he has computed the "community of function" between intelligence tests and arithmetic tests as about 88 per cent, and that between intelligence and reading tests as about 92 per cent. Since a "scant one-tenth" of the tests are utilized in the measurement of difference between intelligence and achievement, he points out the serious hazards of such comparisons. It would certainly appear that unless the apparent differences are very great, the existence of true differences cannot be safely inferred. Conrad [54] has described the conditions which must be met before scores are strictly comparable.

Index of studiousness. Symonds [55] has proposed an *index of studiousness* based upon the differences in ranks of the pupils on intelligence and achievement tests. He suggests two ways of making the comparison, the simplest based on the relative rank in class on the two types of tests. These differences are then added *algebraically* to 100 in order to avoid negative numbers. For example, a pupil ranks fifth in intelligence but stands second in achievement in his class. The difference of 3 is added to 100, giving him an index of studiousness of 103. A second pupil ranks third in intelligence but falls to tenth place in achievement. This is a negative difference of —7, which, when added algebraically to 100, gives an index of studiousness of 93. All pupils who have the same rank on both tests have indices of 100. In an average class of around 35 pupils, an index of 90 to 95 would justify special inquiry, and an index below 90 would give very strong evidence of maladjustment. But it should be noted that in addition to the inadequacies of tests for revealing differences, as pointed out by Kelley, such a system based on relative ranks in the class is limited by the size of the class. Take the case of the boy who reported to his father that he stood second in his class, which sounded all right, until the father discovered there were only two in the class!

---

[53] Truman Lee Kelley, *Interpretation of Educational Measurements,* page 208. Yonkers: World Book Company, 1927.

[54] H. S. Conrad, "Comparable Measures," in *Encyclopedia of Educational Research,* edited by Walter S. Monroe, pages 340–344. New York: The Macmillan Company, 1941.

[55] Percival M. Symonds, *Measurement in Secondary Education,* pages 521–525. New York: The Macmillan Company, 1927.

The accomplishment quotient. In 1920 Franzen[56] suggested the *accomplishment quotient,* abbreviated AQ, for the same purpose. This is the ratio between EA and MA, or between EQ and IQ. The simplest formula is $AQ = EA \div MA$. A quotient of 100 is considered the goal. For example, if a pupil whose MA is 10–0 has an EA of 9–2, his AQ is $110 \div 120$, or 92. In like manner, a second pupil might have the same EA, 9–2, but an MA of only 8–3. His AQ would be $110 \div 99$, or 111. The interpretation of the first case, 92, is that the pupil is not living fully up to his capacity, which seems reasonable enough, human nature being what it is. But the interpretation of the second case, 111, is rather absurd, since it appears to imply that this pupil has exceeded what he is capable of doing by 11 per cent! A far more probable explanation is that the quotient is due to inaccuracies in the tests, and that in this case the errors in the achievement score were in the direction of making it too high, whereas the errors in the intelligence score were in the direction of making it too low. The resulting quotient has added these errors. If the errors, whether due to chance or otherwise, had been in the same direction, they would have tended to offset each other. One reason the use of IQ and EQ involves less risk is that CA, the divisor, is free from error. Studies by Cureton,[57] Haggerty,[58] Tsao,[59] and others have brought the AQ and similar measures into general disrepute.

All things considered, it is probably better to restrict the use of AQ and similar techniques to the measurement of groups, rather than to the measurement of individuals. For example, one might compare the AQ's of two or more classes by dividing the median EA of each by its median MA. Better still, if a teacher has 40 pupils, she might divide them into four groups of 10 each according to their intelligence, and then compute the median AQ for each of the four groups. What she would probably discover is that the average AQ's vary *inversely* with the average intelligence of the groups. In other words, *there is a rather general tendency for the pupils in any grade, who are most retarded educationally from the standpoint of their capacity, to be those of highest intelligence.* Conversely, *there is a rather general tendency for the pupils in any grade, who are most accelerated educationally from the standpoint of their capacity, to be those of lowest intelligence.* Used in this way, the AQ technique would be for the teacher a valuable means of self-analysis. What is revealed should have a most salutary effect on the teacher. It is a tragic fact that praise and blame are not infrequently given to the wrong individuals. Figuratively speaking, the teacher often pats the wrong group on the back, and knocks the wrong group over the head.

---

[56] Raymond Franzen, "The Accomplishment Quotient," *Teachers College Record,* 21: 432–440, November, 1920.

[57] Edward E. Cureton, The Accomplishment Quotient Technic," *Journal of Experimental Education,* 5: 315–326, March, 1937.

[58] Lida Harmar Haggerty, "An Evaluation of the Accomplishment Quotient: A Four Year Study at the Junior High School Level," *Journal of Experimental Education,* 10: 78–89, September, 1941.

[59] Fei Tsao, "Is the AQ or F Score the Last Word in Determining Individual Effort?", *Journal of Educational Psychology,* 34: 513–526, December, 1943.

Combining intelligence and achievement scores. Several proposals have been made for combining scores on intelligence and achievement tests, usually for purposes of pupil classification. One of the simplest of these proposals is to average the pupil's rank on the two tests. More refined methods involve the use of some common denominator, such as the percentile score or standard score. One publication [60] suggests the use of *promotion age* and *promotion quotient* as a basis of classification for instructional purposes. Promotion age (PrA) is the average EA and MA. In this average the two ages may be weighted equally or unequally, whichever seems best for the data in hand. Then the promotion quotient (PrQ) is the PrA ÷ CA. On the face of it, such practice appears to be averaging things as unlike as cattle and horses. But, if Kelley's point regarding the great community of function between intelligence and achievement tests is well taken, the practice would appear to be justified on theoretical grounds. And if it provides a better basis for grouping pupils, as appears often to be the case, it has abundantly justified itself in practice.

F. Use of Norms in Interpreting Personality Scores

As a rule, character and personality measurements do not attempt elaborate systems of norms. At best, such norms as exist are regarded by the authors as provisional. The problem is inherently more difficult than that presented by either intelligence or achievement tests, for which we have seen that the norms are far from adequate. Indeed, the very essence of personality is its uniqueness. It is here that the good judgment and common sense of the teacher are most important.

---

[60] *Supervisors Manual* for the Metropolitan Achievement Tests, pages 38–39. Yonkers: World Book Company, 1933.

Terman[61] strongly questions the possibility, or desirability, of establishing norms for evaluating or adjusting personalities. He says:

> The psychologist stands aghast at the self-assurance with which the professional school counselors in America diagnose the personality faults of little children and at the boldness with which they undertake the delicate task of adjustment. . . . The student of genius who is familiar with the motivating influences that have their origin in quirks of childhood personality shudders to think what the result would have been if school counselors had had a chance to "adjust" the personalities of the budding geniuses of history. One can imagine them, freed from all their peculiarities and complexes, adjusted to the world as it was and becoming indistinguishable from the common herd.

On the same point Poffenberger[62] quotes with approval this statement from Burbank,[63] growing out of a lifelong study of plant life:

> One of the greatest fallacies of near science and of amateurs in Nature's school is the belief that only from the normal can we get our best development and results. As a matter of fact, Nature shows us again and again that it is from abnormalities that some of our most valuable and beautiful plants arise. . . . From that weak, or abnormal plant—that genius plant—may come the very characteristics that we are looking for, and our only problem is to nurse it physically and keep it strong to pass on its overload of spiritual or esthetic essences to its children.

Probably the professional educator could hardly do better than to accept wholeheartedly the motto of the founder of the eugenics movement in England, which was "Treasure your exceptions." Those who deviate most widely from the average deserve special consideration. It is from this group that geniuses are recruited as well as social misfits of all types. It is socially undesirable, as well as psychologically impossible, to make everybody alike.

A distinguished authority on mental hygiene gives this wholesome comment:[64]

> The adjuration to be "normal" seems shockingly repellent to me; I see neither hope nor comfort in sinking to that low level. I think it is ignorance which makes people think of abnormality only with horror, and allows them to remain undismayed at the proximity of "normal" to average and mediocre. For surely anyone

---

[61] Lewis M. Terman, "The Measurement of Personality," *Science,* 80: 605–608, December 28, 1934.

[62] Albert T. Poffenberger, "Psychology and Life," *Psychological Review,* 43: 30, January, 1936.

[63] Luther Burbank and Wilbur Hall, *The Harvest of the Years,* page 273. Boston: Houghton Mifflin Company, 1927.

[64] Karl A. Menninger, *The Human Mind,* page ix. New York: Alfred A. Knopf, Inc., 1930.

who achieves anything is, *a priori,* abnormal; this includes, not only the geniuses, but the presidents, the leaders, and the great entertainers. I presume most of the people in *Who's Who in America* would resent being called normal.

SELECTED READINGS FOR FURTHER STUDY

Boynton, Paul L., *Intelligence: Its Manifestations and Measurement.* New York: D. Appleton-Century Company, 1933. Chapters VIII and IX.

Crawford, John R., "Age and Progress Factors in Test Norms," *University of Iowa Studies in Education,* 9: 1–39, June 15, 1934.

Flanagan, John C., *The Cooperative Achievement Tests, A Bulletin Reporting the Basic Principles and Procedures Used in the Development of Their System of Scaled Scores.* New York: The Cooperative Test Service, 1939. 41 pages.

Freeman, Frank N., *Mental Tests* (Revised Edition). Boston: Houghton Mifflin Company, 1939. Chapters XI and XV.

Greene, Harry A., Jorgensen, Albert N., and Gerberich, J. Raymond, *Measurement and Evaluation in the Elementary School.* New York: Longmans, Green & Company, 1942. Chapters V and XXIII.

Kelley, Truman Lee, *Interpretation of Educational Measurements.* Yonkers: World Book Company, 1927. Chapters II, III, and IV.

McNemar, Quinn, *The Revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin Company, 1942. Chapter XI.

Odell, C. W., *Educational Measurement in High School.* New York: D. Appleton-Century Company, 1930. Chapter XVIII.

Remmers, H. H., and Gage, N. L., *Educational Measurement and Evaluation.* New York: Harper & Brothers, 1943. Chapter XXII.

Stoddard, George D., and others, "Intelligence: Its Nature and Nurture," *Thirty-Ninth Yearbook of the National Society for the Study of Education.* Bloomington, Illinois: Public School Publishing Company, 1940. Part I and Part II.

Stoddard, George D., *The Meaning of Intelligence.* The Macmillan Company, 1943. Chapters IV, XIII, and XV.

Van Omer, Edward B., and Williams, Clarence O., *Elementary Statistics for Students of Education and Psychology.* New York: Longmans, Green & Company, 1940. Chapter VIII.

PART IV

MEASUREMENT IN INSTRUCTION

CHAPTER XI

Motivation

A. The Problem of Motivation

Importance of motivation. It is generally recognized in ordinary experience that motivation occupies an important place in human affairs. Such familiar proverbs as "You can lead a horse to water but you can't make him drink" and "It is hard to teach an old dog new tricks" assign to motivation a key position. The horse does not drink for the simple reason that he does not *want* to drink, and the old dog's poor performance is due not so much to lack of ability as to the fact that he has become too well satisfied with the tricks he already knows. In like manner, every teacher of experience has seen pupils of mediocre capacity succeed because of interest and enthusiasm, while others of more promise have failed utterly because of lack of it. With these observations growing out of ordinary experience the views of psychologists and other keen students of education are in complete accord. Book,[1] for example, asserts that "motivation is the central factor in every learning process." From a careful study of high-school pupils Turney [2] came to the conclusion that the two major factors in school achievement are intelligence and motivation, and that the latter is the more important. Morrison makes the same point in explaining poor performance as due to a lack not of IQ but of *want to.* A "definite and authoritative" consensus of American psychologists agreed that the "most essential thing to know about human nature is *motivation,* the forces, drives, motives that turn the human machinery—that make us think, feel, and particularly act as we do." [3] A similar view has been expressed by Einstein, one of the most eminent of modern scientists: [4]

---

[1] William F. Book, *Economy and Technique of Learning,* page 311. New York: D. C. Heath and Company, 1932.

[2] Austin H. Turney, "Intelligence, Motivation, and Achievement," *Journal of Educational Psychology,* 22: 426–434, September, 1931.

[3] Daniel Starch, Hazel M. Stanton, and Wilhelmine Koerth, *Psychology in Education,* pages 29–31. New York: D. Appleton-Century Company, Inc., 1941. Reprinted by permission.

[4] Albert Einstein, "Some Thoughts Concerning Education," *School and Society,* 44: 590, November 7, 1936.

> But behind every achievement exists the motivation which is at the foundation of it and which in turn is strengthened and nourished by the accomplishment of the undertaking. Here there are the greatest differences and they are of greatest importance to the educational value of the school.

Meaning of motivation. The term *motivation* is very inclusive. Literally it means *causing movement.* Young uses the term to refer to "all conditions which arouse and regulate the behavior of organisms." [5] Irwin, in substantial agreement, says that the term "includes those factors which in an individual and in the situation determine the nature of his acts." [6] A convenient grouping of motives into two major classes is suggested—internal or organic, and external or environmental. In recent years the term *drive,* or *urge,* has been used for the former, and *goal,* or *incentive,* for the latter. But in the final analysis, motivation, though in some instances externally initiated, always functions internally. Hunger, thirst, and sex, as well as interests, attitudes, wants, desires, and temporary mental sets, are examples of drives. Incentives may be negative, as are pain or punishment, or positive, as are rewards in a multitude of forms. A further distinction is often made between motives which are natural or intrinsic, such as a child's interest in play or the movies, and those which are artificial or extrinsic, such as prizes, marks, grades, credits, and honor rolls.

Relation of measurement to motivation. Measurement is related in at least two ways to motivation. In the first place, there is the problem of the measurement of motivation itself. It is often important to know the differences among individuals in the strength of various motives, the comparative strength of the same motive under varying conditions, and the strength of a given motive in comparison with other motives in the same individual. As the development of wholesome attitudes and interests is an objective of modern education, it is just as necessary to know how to measure it as to know how to measure any other objective. While much valuable work has been done by Moss, Warden, and others in the measurement of animal drives, up to the present no convenient technique has been devised for measuring human motives in any precise manner.[7] The measurement of motivation in education is, then, a problem for the future. Thorndike, one of the leading students of motivation, ventures the guess that differences in native wants are of the order of differences in original capacities of intellect or skill, while differences in acquired wants are probably even greater than the differences in acquired abilities.

---

[5] Paul Thomas Young, *The Motivation of Behavior,* page 45. New York: John Wiley & Sons, Inc., 1936.

[6] M. E. Irwin, in *Review of Educational Research,* 6: 300, June, 1936.

[7] Edward L. Thorndike, *The Psychology of Wants, Interests, and Attitudes,* pages 14–15. New York: D. Appleton-Century Company, Inc., 1935.

In the second place, there is the problem of the relation of the measurement of educational capacity and achievement to the motivation of learning and teaching. Since teaching and learning are two aspects of the same process, it is reasonable to expect that measurement will be intimately related to both. Some of the more important relationships will be considered in the next two sections.

B. The Relation of Measurement to Motivation in Teaching

Purpose of the teacher and measurement. An obvious relationship of measurement and motivation in teaching arises from the fact that the purpose of the teacher determines the type of measurement used. Whether, for example, the teacher gives many tests or few, long tests or short, informal tests or standardized, survey tests or diagnostic, depends upon his purpose. Since not all tests serve the same purpose equally well, as has been pointed out, the choice of the measuring instrument becomes a matter of primary importance. This problem has already been considered at some length in Chapter III. Certain points to be raised later in this chapter will also have a bearing upon it.

Teaching emphasis and measurement. The proper teaching emphasis is determined by the results of measurement. Measurement directly demonstrates the quality of the pupil's learning, but it also indirectly reflects the quality of the teacher's teaching. In the light of measured results conscientious teachers attempt as far as possible to correct weaknesses in past teaching and to prevent their recurrence in future teaching. Messenger [8] studied the "influence of the Iowa Academic Testing Program in relation to the teaching of English mechanics in an Iowa high school" and found that the effect had been to "motivate teachers to greater effort," with the allotment of more time to the subject and the use of more drill material. One of the chief values of measurement may well be its motivating effect upon the teacher.

Taba, an able defender of progressive education, realizes this relationship and speaks regretfully of the "formidable and serious handicap" of the progressive schools which is owing to the "lack of forms of testing that are in harmony with their aims and adequate to their purposes." Then she adds these significant words: [9]

> After all, one teaches only what one, in some way or another, is able to evaluate as an outcome of that teaching. If we are unable to evaluate the growth of in-

---

[8] Unpublished master's thesis, University of Iowa, 1934.

[9] Hilda Taba, *The Dynamics of Education,* page 185. New York: Harcourt, Brace and Company, Inc., 1932.

tegrations and meanings and ways of behavior, we are unable even to form an adequate notion of them, still less to guide the process of learning in these terms.

Attention should be called to the fact that Taba's recognition of the limitations of existing measurement does not blind her to the important motivating influence of measurement, whether good or bad, and to the urgent necessity for continuous improvement of measuring instruments.

There is, of course, some danger that the content of the examination may exert too great an influence over the teaching emphasis and curriculum content. It has often been alleged that something like this has happened in the case of the New York Regents and the College Board Examinations, where an important effect of such examinations has been to turn secondary schools into "cramming" schools pointing toward the probable content of such examinations as revealed by past examinations by the same agencies. Insofar as this is true, it represents unfortunate and unwarranted control over teaching procedures, which not only defeats the purpose of the examination but places an obstacle in the way of education itself. Such practice fails to take into account the sampling nature of all tests. Any attempt to drill pupils in advance specifically upon the items of the test tends to narrow teaching to the scope of the testing, so that the two tend to become synonymous rather than one a random sampling of the other. This should be clearly understood, and every reasonable precaution should be taken to insure that state-wide testing programs and other forms of academic competition do not degenerate into mere monotonous drill exercises of an especially narrow and vicious type.

A few studies have been made of the effect of exempting pupils from final examinations. An early study by Anderson[10] indicated that exempting high-school pupils who reached a minimum standard from final examinations had "played havoc with teachers' grades," while the actual performance of the pupils had shown "no appreciable increase." Apparently the motivation had been in the wrong place. That such a disastrous result need not occur is indicated by a later study of the same problem reported by White,[11] who found that the general distribution of marks in his school had changed very little under the exemption system. Even here, however, was found a "decided dip in the distributions immediately below the exemption point of 85 per cent and a corresponding rise just above it." Schools employing an exemption system should remain constantly on guard lest its effect be more to stimulate the teachers to *give* high marks than the pupils to *earn* them.

---

[10] C. J. Anderson, "Is the Exemption System Worth While?" *School and Society,* 3: 357–360, March 4, 1916.

[11] Clyde W. White, "The Effects of Exemptions from Semester Examinations on the Distribution of School Marks," *School Review,* 39: 293–299, April, 1931.

C. The Relation of Measurement to Motivation in Learning

Close as is the relationship of measurement to the motivation of teaching, the relationship is even closer to the motivation of learning. Elsewhere the writer [12] has stated the problem as follows:

> Behind the *act* of learning is the *capacity* to learn, and back of the *capacity* is the *motive* to learn—the desire, urge, impulse, drive, or something, that makes the creature *want* to learn, that pushes him out to meet his environment. One of the reasons why most of the correlations of mental capacity with actual achievement, in school and out, have been disappointingly low, has been that students of real ability have not felt a proper urge to work, while those of mediocre talent have frequently possessed the urge to achieve.
>
> If we are ever to be successful in our efforts to predict achievement, therefore, we must not be content with merely analyzing the learning process, understanding the mechanism of learning, its structure and laws of operation, nor with merely exploring the height and range of human possibilities; but we must also find out about the dynamic aspects of human nature. We must discover not only *how* the mind works, but *why* it works when it does and the way it does.[13]

The foregoing statement is an introduction to a report of an experimental attack on one phase of motivation. The study will be briefly presented here as an illustration of one type of psychological experiment concerned with this important problem. Afterwards some critical comments upon this and other similar experiments will be given.

An experiment in motivation. The problem was to determine the influence of a knowledge of results upon the achievement of 59 college students in a simple act of motor skill, making tally marks (~~////~~). The procedure was as follows: Upon the basis of an initial practice period three equivalent groups were formed. One of these, the control, had *no knowledge* of its progress, throughout the first ten practice periods of one minute each. During this time one experimental group had *full knowledge* of results, and the other had *partial knowledge.* At the beginning of each practice period after the first, each pupil in the group with full knowledge was shown his paper of the preceding day, with scores and corrections indicated. A distribution of scores for this group was placed on the board, and each student was urged to watch his daily progress, both relative and absolute. In the experimental group with partial knowledge of results, each student was told whether he was above or below the average of the group, but that was all. At the end of ten practice periods conditions were reversed; the group which had had no knowledge of results was then given full knowledge for two additional periods, and the other two groups were given no knowledge for these two periods.

---

[12] Clay Campbell Ross, "An Experiment in Motivation," *Journal of Educational Psychology,* 18: 337–346, May, 1927.

[13] *Ibid.,* page 337.

The results are shown in Figure 30 on page 275. On the whole, they seemed to justify the conclusion that "the addition of a single other motivating factor, knowledge of results, is sufficient to give the pupils with such knowledge a distinct superiority over the others, and the degree of superiority is roughly proportional to the amount of information possessed."

Limitations of experiments on motivation. The experiment just summarized illustrates several of the weaknesses of the experimental work so far reported on motivation and, for that matter, on other phases of learning as well. These may be conveniently grouped under three headings: factors studied, subjects used, and conclusions drawn.

In the first place, the factors so far studied leave much to be desired. Most of the studies involved are concerned with highly artificial and often trivial tasks. Take the above experiment as an example. It is highly improbable that students will show much enthusiasm over making, as rapidly as possible, groups of four vertical lines crossed horizontally with a fifth. Tallying, to be significant to most persons, would have to be employed as a record of some athletic contest or other situation in which it is a means to an end and not an end in itself. A survey of the literature reveals that a large percentage of motivation studies have been concerned with making legible *a*'s, canceling numbers or letters, assigning a number to a dictated word, learning trivial facts or actual misinformation, running mazes, and the like. What we need to know is how people behave under actual school conditions. Even when arithmetic and other school materials are employed, the experiments are rarely carried on long enough for the novelty of the task to wear off and for the experimental factor to operate under reasonably normal conditions. The total learning time is frequently less than one hour, and often it is only ten to fifteen minutes, as in the above laboratory experiment. To be most helpful in guiding teachers in the day-by-day conduct of their classes these experimental factors must be continued for at least several weeks.

In the second place, the choice of subjects for the experiment has usually been rather unfortunate. Many of the laboratory studies of the effects of rewards and punishments have been limited entirely to animals. No matter how thorough a believer one might be in evolution, one must certainly see that the behavior of rats in a maze or cats in a puzzle box might very well be different in essential respects from that of school children facing the intricacies of a foreign language or learning to manipulate the abstract symbols of algebra. Even when human subjects have been used, as they have been in many experiments, they have usually been adults, often graduate students of psychology. Often, also, the number of subjects has been small, with a poorly equated control group, or with none at all. To be most convincing as a guide for school practice these experiments must be performed with children of approximately the same age and type as those found in the schools where the results are applied. If our studies of either maturation or learning are to be relied upon, the child is not merely a miniature adult. And if the numerous studies of individual differences from the time of Galton to the present have established anything, it is that what is true of one person is not necessarily true of another. Experiments based on a handful of subjects, therefore, must be accepted with considerable discount.

In the third place, the conclusions drawn have often gone far beyond the experimental facts available. This is mainly the result of the two limitations already mentioned. If experimenters would be content to draw conclusions from, and to make applications to, the same or closely similar subjects and to the same or closely similar tasks, no harm would be done. But this is rarely the case. To generalize from one age level to another is risky even when the task is the same, but it is particularly hazardous when the activity itself is different. Yet this very thing is commonly done. A meaningless and often trivial act is performed by adults under the highly artificial conditions of the laboratory. Then the results are applied without qualification to the meaningful learning of children under the actual conditions of the schoolroom. That this procedure is wholly unwarranted will appear from an experiment by the writer to be reported later in the chapter. But to generalize from the behavior of a rat in a psychological laboratory to that of a child in an ordinary classroom is little less than foolhardy.

Types of motivation experiments. From what has just been said, it may appear that the writer believes all motivation experiments to be worthless. Such is far from his conviction, however. While he does believe that practically all such experiments so far reported have certain weaknesses to which attention should be called, he is convinced that genuine progress has been made and that the way has been opened for further studies to supplement those already in existence. It has been definitely shown that the problem, although difficult, is susceptible to experimental attack. Furthermore, experimental evidence already available is sufficient to provide at least tentative answers to two important questions:

1. What is the relation of measurement to the *amount* and *quality* of learning?
2. What is the relation of measurement to the *type* of learning, or to the *learning procedure* followed by the student?

The experimental literature will now be reviewed in relation to those questions, and certain generalizations growing out of it will be attempted.

I. The Relation of Measurement to the Amount and Quality of Learning

There is considerable experimental evidence regarding the influence upon the *amount* and *quality* of learning of three measurement factors, or groups of factors:

a. The frequency of the tests.
b. The knowledge that a final examination would be given.
c. The knowledge of results or progress in learning.

Attention has been given to the operation of these factors individually, in combination with each other, and in combination with other motivating factors, such as praise and blame, rivalry, and various types of material rewards. Some of the more important of these studies will now be summarized.

Frequency of tests. Practice varies widely regarding the frequency of testing. At one extreme are the teachers who give no written examination of any kind, and at the other extreme are those who give a test of some kind every day. What experimental evidence is there to indicate the proper frequency? Manifestly, whatever advantage may exist in frequent testing cannot be attributed solely to the motivating effects, however, since the additional practice afforded by taking the extra tests must also be considered.

The experimental evidence regarding proper frequency of testing in social studies classes in high school has not been very convincing. Hoglan [14] studied the frequency of testing in American history in one Iowa high school. He found no significant differences among three groups, equated on the bases of intelligence and knowledge of history. One group had daily tests, another had three unannounced tests per week, and a third had only the regular tests at intervals of six weeks. In a similar study in the same subject Camp [15] found a slight (but not a statistically significant) difference between one group tested once or twice a week and another tested once in two or three weeks. It should be noted, however, that the difference would have to be large to be statistically significant where there are only 46 in each group. In an earlier study in community civics Shore[16] found no advantage in giving each day a true-false test of ten items; but a group given an unannounced test two or three times a week did show a statistically significant superiority over a group given only the mid-semester and the final tests. On the whole, the evidence appears to favor slightly the practice of giving a test once or twice a week to classes in the social studies in high school.

---

[14] Unpublished master's thesis, University of Iowa, 1932.
[15] Unpublished master's thesis, University of Iowa, 1931.

Two experiments on the effects of frequent testing of high school biology students have reported conflicting results. Kitch[17] found that the group taught with the aid of self-scored unit tests did significantly better than the group without such tests. Gable[18] compared the merits of three procedures. One group was told that it would be tested each day, another group that it would be given announced unit tests, and a third group understood that it would be tested without notice at irregular intervals. On the whole, the poorest record was by the group taking daily tests, but there was a tendency for the slower pupils to do better when a test was announced which gave time for review.

Three studies have been reported on the motivating effect of frequent testing in high-school physics. Conner[19] found that the use of a well-known series of instructional tests in physics had not resulted in sufficient improvement in learning to justify the time expended. Kugle[20] reported that short daily tests in physics resulted in pupils' having a small superiority over those to whom tests were given only at the ends of units. Kirkpatrick[21] found a distinct advantage, in the 26 high-school physics classes included in his study, in giving an objective test at the beginning of each unit. As each unit covered from one to three days, this meant that tests were given at least twice a week. The pupils had definite knowledge that the test would be given, that it would cover all the important concepts of the unit, and that the final examination would include only points included in these unit tests. The tests were corrected in class and were used as a basis for class discussion and subsequent study. Both experimental and control groups took the same term tests at intervals of six weeks. When the experimental groups were considered as a whole, a statistical difference that is highly significant was found on a test of objective information given at the end of the course, but this superiority had largely disappeared four months later. The testing program was most beneficial to the pupils in the lowest third in mental ability. This suggests that schools attempting to group pupils according to ability may very well consider varying the testing program as well as the curriculum and teaching methods.

---

[16] Unpublished master's thesis, University of Iowa, 1925.

[17] Loran V. Kitch, unpublished master's thesis, University of Southern California, 1932.

[18] Sister Felicita Gable, *The Effect of Two Contrasting Forms of Testing Upon Learning*. Baltimore: Johns Hopkins Press, 1936.

[19] Unpublished master's thesis, University of Iowa, 1932.

[20] Unpublished master's thesis, Pennsylvania State College, 1936.

[21] James Earl Kirkpatrick, "The Motivating Effect of a Specific Type of Testing Program," *University of Iowa Studies in Education,* 9: 41–68, June 15, 1934.

Experiments involving the frequency of testing have been most numerous and, on the whole, most convincing on the college level. A serious limitation, however, is that they have been largely restricted to classes in general and educational psychology. Jones,[22] in a pioneer study, gave five-minute completion tests, euphoniously called "terminal reviews," at the end of each of 27 lectures in psychology. Eight weeks later the groups so tested made scores on a final examination that were approximately twice as high as those of groups who had had no "terminal reviews." Another study [23] reports advantages to be gained in using weekly objective tests in general psychology. Jones' argument for the "marked advantage in presenting the examination on a given lecture immediately at the close of that lecture rather than at any later time" is as follows: "Examination strengthens connections. But the later an examination is given, after the original lecture, the fewer are the connections which remain to be strengthened."

Both Turney [24] and Keys [25] have found weekly tests in educational psychology better than tests given less frequently. Turney found that, when given weekly tests, a class which was 20 per cent below another class at the beginning was able to equal the achievement of the other class, which had only one short test, in addition to the mid-semester and the final, which both groups had. Keys found that eight weekly tests gave a 12 per cent advantage over the same items given to equivalent groups in the form of two monthly tests. However, on an unannounced examination covering the same material given five weeks later, this advantage had been reduced to 7 per cent. When the regular final examination came, after the additional two weeks, the achievement of the experimental and control groups was practically identical. What the effect of the weekly testing was after still larger intervals is, unfortunately, unknown. This opens up an important field for further study.

---

[22] Harold E. Jones, "Experimental Studies of College Teaching," *Archives of Psychology,* 68: 36–70, November, 1923.

[23] C. C. Ross and Lyle K. Henry, "The Relation between Frequency of Testing and Progress in Learning Psychology," *Journal of Educational Psychology,* 30: 604–611, November, 1939.

[24] Austin H. Turney, "The Effect of Frequent Short Objective Tests upon the Achievement of College Students in Educational Psychology," *School and Society,* 33: 760–762, June 6, 1931.

[25] Noel Keys, "The Influence on Learning and Retention of Weekly as Opposed to Monthly Tests," *Journal of Educational Psychology,* 25: 427–436, September, 1934.

Johnson [26] compared the effect of written unit tests and the effect of an equal amount of time devoted to oral reviews with 55 pairs of freshman girls in two classes in child development. She found that a statistically significant difference in favor of the tested group had disappeared twelve weeks later. She concluded that "there is as yet no evidence to show that the greater achievement which has been induced by examinations persists after six weeks to three months."

A few studies have reported little or no advantage in weekly tests, even when comparisons were made at the end of the course. For example, weekly tests in general psychology at the University of Minnesota gave negative results.[27] Both Noll [28] and Ross and Henry [29] found a slight superiority in less frequently tested groups in educational psychology. However, Ross and Henry in both general and educational psychology, and Noll in educational psychology, found evidence that the benefit of weekly tests was greatest for the students of low ability. It is evident that there is no one best testing technique which is equally effective under all conditions. Testing methods as well as other teaching procedures must consider the ability of the student as well as the nature of the subject.

Kulp [30] gave the students in a graduate class in educational sociology who were below the median on the mid-semester examination a weekly ten-minute objective test for the next seven weeks. The students above the median were excused from these short tests. On the final examination, "identical in all respects with the seven weekly tests," the superiority of the upper half, which had been about 39 per cent at mid-semester, was reduced to 5 per cent.

---

[26] Bess E. Johnson, "The Effect of Written Examinations on Learning and on Retention of Learning," *Journal of Experimental Education,* 7: 55–62, September, 1938.

[27] A. C. Eurich, H. P. Longstaff, and M. Wilder, *The Effective College Curriculum as Revealed by Examinations,* pages 333–347. Minneapolis: University of Minnesota Press, 1937.

[28] Victor H. Noll, "The Effect of Written Tests upon Achievement in College Classes: An Experiment and a Summary of Evidence," *Journal of Educational Research,* 32: 345–358, January, 1939.

[29] C. C. Ross and Lyle K. Henry, *op. cit.,* pages 609–610.

[30] Daniel H. Kulp, II, "Weekly Tests for Graduate Students?" *School and Society,* 38: 157–159, July 29, 1933.

Pressey[31] reports an interesting variation of this procedure as used by Smeltzer in educational psychology. Both experimental and control classes were given weekly tests. But in the experimental class, to whom the test was given on Thursday of each week, the papers were returned and discussed on Friday. Those who had made unsatisfactory scores were tested again over the same material after a brief review on Monday, while the others were excused. On the final examination the experimental group was approximately 5 per cent above the control group, the advantage being largely with the pupils who were in the lowest fourth of the class and who had taken the retests.

Three of the above experiments attempted to get the students' attitude toward the frequent testing. By means of unsigned questionnaires in three classes, Jones found that 70 per cent of the students approved the "terminal review method." In like manner, Turney discovered an "excellent attitude" toward frequent testing in his experimental group; about 85 per cent thought they had studied more, and over 90 per cent said that they preferred to be in that section and that they felt they had learned more even if they had made no better grade. From "an extensive questionnaire touching some thirty issues of educational theory and practice," given at the opening and repeated near the end of the semester, Keys found: "Without comment by the instructor or knowledge of the experiment in progress, students disclose a strong and growing conviction of the desirability of tests given as frequently as every second, third, or fourth class session." The evidence strongly suggests that students favor frequent testing.

Awareness of final examination. To what extent is the "intention to remember" or "temporal set" a factor in learning? Will the expectation that the material will have to be recalled later influence the amount retained? More specifically, how will the awareness of a final examination affect the progress of learning and of forgetting? One or the other of the following "two rival and mutually exclusive hypotheses," as suggested by Remmers,[32] is apparently true:

1. Exemption from final examinations with its requirement of continuous high-level learning provides better motivation and therefore more permanent learning and integration than does the final examination.
2. The final examination provides the opportunity and at least a part of the stimulation for the better development of certain abilities such as rapid organiza-

---

[31] Sidney L. Pressey, *Psychology and the New Education*, pages 363–366. New York: Harper & Brothers, 1933.

[32] H. H. Remmers and others, "Exemption from College Semester Examinations," *Purdue University Studies in Higher Education*, 23: 11, November, 1933.

tion of a large mass of material; the ability to select crucial data from the large mass of material; to see pertinent relationships, to reason in terms of the subject matter, to apply this reasoning to significant problems, etc.; and in general more effective and permanent learning.

Thisted and Remmers [33] summarize the literature on the general problem, including studies on such dissimilar materials as stories, objects shown, vocabulary, nonsense syllables, photographs, and stylus mazes. They conclude: "It is evident that a condition of expectation of recall, when injected into the initial instructions, has given variable and conflicting results." Their own study, which included 404 psychology students, involved learning Anglo-Saxon vocabulary and the factual content of two articles presented in mimeographed form under ordinary classroom conditions. The control group understood that they were to be tested immediately, and the experimental groups understood that they were to be tested later also, after three days in some cases, after one week and after two weeks in other cases. The experiments tended to establish a somewhat slower drop in the forgetting curve when a temporal set was introduced in order that delayed recall would be required.

While the learning material in the above experiments was not left in the hands of the students, it is reasonably sure that one effect of the "temporal set" was to cause those who expected to have to recall the material later to give a "mental review" of what they could remember, as well as probably to exchange ideas with other students. Under ordinary school conditions one might expect that the effect of reviewing for an expected examination might be larger. However, Remmers [34] later found that exempting students in mathematics and applied mechanics made "relatively little difference in the amount, quality, or permanence of learning, at least as measured by current types of tests and examinations."

Pease [35] reports some interesting studies of the effect of "cramming" on the amount of class materials retained. The first study included several classes—in all, 302 college students and 106 high-school pupils—separated into equivalent groups on the basis of intelligence. A test of 100 objective items was prepared for each class, "covering several months of the usual course work already completed by the students." At a meeting of the class the purpose of the experiment was clearly explained. The experimental group in each class was then dismissed, with instructions to spend at least an hour in review for the examination which was to come at the next meeting of the class. The control group in each class took the examination at once without warning or review. The mean score of the experimental group exceeded that of the control group in each class, the average superiority being 11.1 points on the 100-item test. Without warning, the test was repeated six weeks later. The average lead of the experimental group had been reduced to 6.3 points. But there was still a significant difference for all classes containing as many as fifteen pairs of pupils. However, when one of these classes was retested after an additional six weeks, the lead of the experimental group was reduced by about half and was not then reliable. After twelve weeks the lead in another class had been reduced from 17.07 points to 2.7 points, which was not a significant difference. These results had been produced by an amount of "cramming" by the experimental groups that represented about one and one half hours, on the average. It appeared probable that the time so spent yielded returns that averaged higher than the same amount of time spent either in class attendance or in regular preparation outside. Pease concludes that "from the standpoint of the student, it pays to cram."

---

[33] M. N. Thisted and H. H. Remmers, "The Effect of Temporal Set on Learning," *Journal of Applied Psychology,* 16: 257–268, June, 1932.

[34] H. H. Remmers and others, *op. cit.,* page 52.

[35] Glenn R. Pease, "Should Teachers Give Warning of Tests and Examinations?" *Journal of Educational Psychology,* 21: 273–277, April, 1930.

Tyler and Chalmers [36] studied the effect on test results of warning junior-high-school pupils that they would have a unit test in general science on the following day. The test scores of pupils so warned were compared with comparable pupils who had no specific warning, although they were all aware that it was customary to have a test at the end of each unit, usually with the time announced at least two days in advance. All of the obtained differences favored the warned groups but by margins below the level of statistical significance. Six weeks later, when the tests were repeated, the differences had practically disappeared. The authors questioned whether junior-high-school pupils are really motivated to study for unit tests even when announced, or know how to study effectively when they try. To be effective, motivation has to be intelligently directed.

White [37] conducted an experiment that bears directly upon the effect of exemption from a final examination. Three classes in general psychology which met once a week for seventeen weeks were divided, according to chance, into experimental and control groups. At each weekly class meeting both groups were given a "comprehensive, mimeographed true-false test covering the chapters studied for the period." From the outset the control groups understood that their marks in the course would be based solely upon these weekly tests, while the experimental groups understood that they were to have a final examination that would count 50 per cent toward their course marks. At the class meeting following each test the corrected papers were returned to all students, and they were allowed to keep the papers. The experimental groups were urged to preserve these test papers for further study, as the final examination would contain exactly the same items. At the end of the seventeen weeks the final examination was given to both groups, the hearty co-operation of the students in the control groups being asked in order to determine the value of the experiment. The difference between the groups was 51.2 per cent, the experimental group having *gained* 31.6 per cent and the control group having *lost* 19.6 per cent. Even more convincing was the equal superiority of the experimental group on a completion test "with which they were wholly unfamiliar."

---

[36] F. T. Tyler and T. M. Chalmers, "The Effect on Scores of Warning Junior High School Pupils of Coming Tests," *Journal of Educational Research,* 37: 290–296, December, 1943.

[37] Hubert B. White, "Testing as an Aid to Learning," *Educational Administration and Supervision,* 18: 41–46, January, 1932.

Knowledge of test scores. What is the effect upon the course of learning of the knowledge of progress, afforded by test scores or by other means? The answer to this question has been sought many times in the psychological laboratory, with practically unanimous results. Psychologists are in substantial agreement with the conclusion of an early study[38] that "the addition of a single other motivating factor, namely, knowledge of results, is sufficient to give the pupils with such knowledge a distinct superiority over the others, and the degree of superiority is roughly proportional to the amount of information possessed." However, as has been pointed out earlier in the chapter, experiments conducted in the classroom are far more convincing. We shall now take a look at what they have shown.

One of the earliest and most comprehensive of these studies conducted under actual schoolroom conditions was that of Panlasigui.[39] The findings were based on 358 pairs of pupils in fourth-grade arithmetic in ten cities. The practice material consisted of fifteen minutes' drill in examples of the mixed type of fundamentals once a week for twenty weeks. As all pupils scored their papers after each drill, it can be seen that each pupil knew his *achievement for the day,* although this knowledge must be related to previous records in order to be a *knowledge of progress,* strictly speaking. In the experimental classes the idea of progress was stressed, progress charts for both the individual and the class being kept in a conspicuous place. The teachers of the control classes, on the other hand, were instructed as follows: "Please keep very much out of the class discussion any reference about how much pupils are scoring." The comparison appeared to be, then, between experimental classes with somewhat *more* stress on progress, and control classes with somewhat *less* stress on progress, than is customary to the ordinary teacher. On a comprehensive test the mean of the experimental classes exceeded that of the control classes by 11.34, with a probable error of the difference of 2.88. A detailed examination of the results reveals the fact that this superiority is most in evidence in the highest quarter and practically non-existent in the lowest quarter. "The beneficial effect of awareness of success, then, was substantially in direct proportion to the amount of success available for motivation." This is also true of the drill periods themselves, where the accuracy standards of the highest fourth of the experimental group exceeded those of their controls eight times out of eleven, whereas the lowest fourth of the experimental groups fell behind their controls on every drill. This experiment seems to have established rather definitely two important points:

---

[38] See page 319 for a brief description.

[39] Isidoro Panlasigui, *The Effect of Awareness of Success on Skill in Arithmetic,* unpublished doctor's dissertation. Iowa City: University of Iowa, 1928. For a brief account, see: *Twenty-Ninth Yearbook of The National Society for the Study of Education,* pages 611–619. Bloomington, Illinois: Public School Publishing Company, 1930.

1. A knowledge of progress in learning under classroom conditions is likely to have much less effect than that under laboratory conditions.
2. A knowledge of progress is likely to be more beneficial to good students than to poor.

So far as the writer is aware, all studies since reported have without exception confirmed the first point, but most of them, unfortunately, have not been analyzed with respect to the second point.

Forlano[40] conducted a comprehensive series of experiments in grades four to eight, inclusive, involving in all 1,294 pupils, and touching upon various aspects of the problem of the effect on learning of a knowledge of results. The experimenter emphasizes the fact that these studies were made "in the normal classroom situation . . . as far as possible as a part of the daily school routine." He attempted to determine whether giving a knowledge of results *immediately* after the word had been spelled or an arithmetic fact had been studied was more effective than when a knowledge of results was *withheld* until an entire column of 20 or 24 items had been attempted. In other words, if one may use the analogy of the target range, Forlano was interested in finding out, so to speak, whether it was better to tell the marksman his score after each shot or to wait until he had fired a series of 20 shots. The author's conclusion is as follows: [41]

---

[40] George Forlano, *School Learning with Various Methods of Practice and Rewards*, pages 55–114. New York: Bureau of Publications, Teachers College, Columbia University, 1936.

> The results of our experiments show that there is a tendency for learning during which the learner ostensibly receives immediate knowledge of results to be less efficient than learning in which knowledge of results is delayed. In general, it may be said that this superiority of the "delayed knowledge of results" method does not always approach statistical certainty.

Even this modest conclusion, the author suggests, is limited by the fact that the methods employed "may not be 'pure' methods of what they purport to involve," and that the apparent superiority of the delayed procedure may be due to other causes. In any event, since the period of delay never exceeded five minutes, little light is shed upon the ordinary school situation, where the tests follow learning after an interval ranging from a day to a year or longer. It is, therefore, to be hoped that other studies will follow, comparing the effects of longer delays.

Brown [42] reports an experiment in arithmetic in grades 5A and 7A. Both his procedure and his conclusions differed somewhat from those of Panlasigui. In grade 7A, Brown selected his experimental and control grades, which were only roughly equivalent, on the basis of an intelligence test, and in grade 5A on the basis of estimates of intelligence and achievement. The groups were reversed at the end of the first period of ten days. The drill period was eight to ten minutes daily. While the differences, on the whole, favored the experimental group, they were not very impressive. An examination of the individual drill periods reveals the fact that the progress from day to day in all groups was irregular and somewhat inconsistent, and that the differences between experimental and control groups were generally less on the tenth day than on the first. There was some evidence in Brown's study that the incentive was somewhat more effective with boys than with girls, but the outstanding fact was the remarkably small amount of influence, taken from any point of view, of a knowledge of progress in the classroom as compared with the laboratory.

Deputy [43] conducted in a state university a carefully planned experiment with three groups of students, of approximately equal intelligence, in freshman philosophy, which met twice a week. For six weeks during the first half of the semester the first ten minutes of each class meeting of the control group were devoted to an oral review of the preceding lesson. One of the experimental groups had a ten-minute objective test covering the same material, and the other experimental group had the same items in a twenty-minute test given once a week. Beginning at the middle of the semester, the group which had served as a control was given the ten-minute test at each class meeting, while the other two groups had only the oral reviews. The scores for the experimental groups were put on the board following each test, and each student was urged to keep a record of his progress. Only one of the three comparisons between the ten-minute written test and the ten-minute oral review showed the former to be superior by a statistically significant amount. This fact the author ascribed to a particularly favorable attitude on the part of the students. The experimental group which excelled happened to be slightly the most intelligent of the three, and also showed itself superior to the group which took the twenty-minute test once a week. Deputy's most significant conclusion was: "Considerable precaution should be taken in applying principles, derived from laboratory and other non-classroom situations, to work in school subjects."

---

[41] *Ibid.*, page 99.

[42] Francis J. Brown, "Knowledge of Results as an Incentive in Schoolroom Practice," *Journal of Educational Psychology*, 23: 532–552, October, 1932.

[43] E. C. Deputy, "Knowledge of Success as a Motivating Influence in College Work," *Journal of Educational Research*, 20: 327–334, December, 1929.

Two years later the author [44] began a series of experiments which were to force him to this same conclusion. Attention has already been called to the earlier laboratory experiment [45] which had appeared convincing not only to the author at the time but also to many readers since that time, judging from the writers on educational psychology who have quoted it with approval.

Upon the basis of a comprehensive examination given at the end of the first unit in a class in tests and measurements, a large class was divided into four substantially equivalent groups. A regular class test was given to all students once a week for the next two months. At the next class meeting following a test, a distribution for the entire class was put on the blackboard and a brief discussion given of each item missed by any considerable number. But the four groups were given different degrees of information as to progress. One group was given *no knowledge* whatsoever as to its scores. A second group was given *vague knowledge*, each student being told merely that his score was "good," "fair," or "poor." A third group was given *partial knowledge*, each student being told his point score but not allowed to see his paper. The fourth group was given *full knowledge*, each student being shown his paper at the close of the class and allowed to ask any questions he wished to ask regarding it.

---

[44] C. C. Ross, "The Influence upon Achievement of a Knowledge of Progress," *Journal of Educational Psychology*, 24: 609–619, November, 1933.

[45] See pages 319 to 321.

Figure 38 shows the results for the four groups in the form of cumulative scores, week by week, for the first eight weeks and for the last four weeks, when the groups were reversed. Nowhere was there a statistically significant difference between any of the groups.

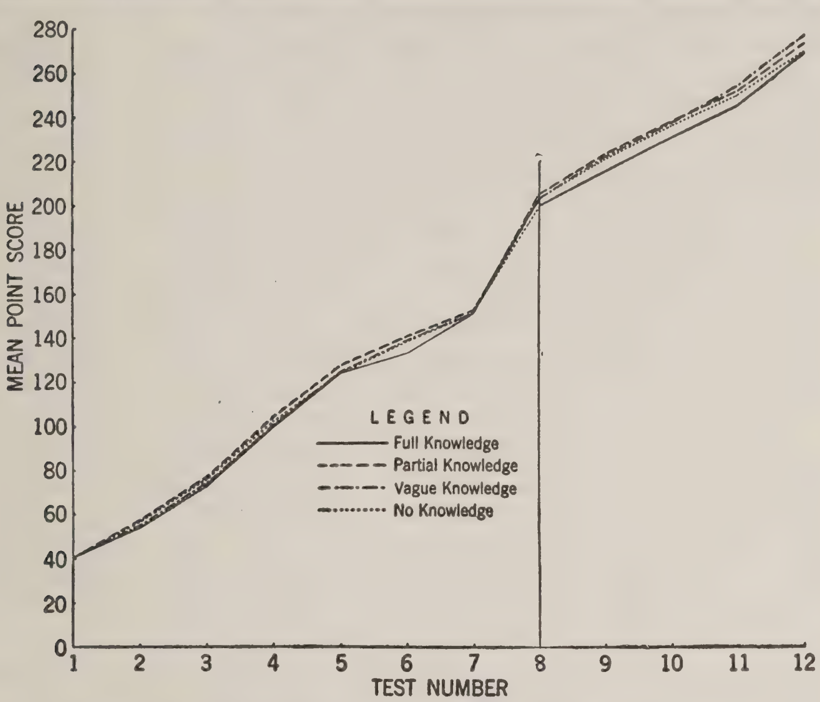


Figure 38. The Influence of a Knowledge of Progress upon Achievement in a College Class.

The experiment was repeated with two other classes in the same subject and with one in a different subject. Not content with this, the writer persuaded a colleague in another department to do the same experiment. In all, the groups involved more than 50 tests and about 300 students, and not once did there appear a difference, favoring the group with full knowledge of progress, that meets the minimum requirement for statistical significance.

Two conclusions seem reasonably certain. The first, directly in line with that of Deputy, is as follows:[46]

---

[46] C. C. Ross, "A Needed Emphasis in Psychological Research," *Psychological Review,* 43: 197–206, May, 1936.

> The Gestalt of the laboratory situation is so different from that of the life situation outside that it is hazardous to generalize from one to the other. One can never be certain what the outcome of a laboratory experiment will be when applied to the classroom situation until it has actually been tried out in that situation.

The second conclusion is that most, if not all, experiments relating to knowledge of results in learning have involved another erroneous assumption: namely, that because students were not told their individual scores, they "had *no* knowledge of progress." Certainly they had their subjective impressions. To test out the accuracy of these impressions the author requested the students in the "no knowledge" group to estimate the scores they thought they had made when they turned in their papers at the close of the tests. The median coefficient of correlation between these estimates and the actual scores was .71. Manifestly, then, such studies involve a comparison of *two kinds of knowledge,* subjective and objective. Moreover, there was a tendency for the poorer students to overestimate their scores. In such cases the illusion of success may very well have proved more stimulating than the reality of failure.

Knowledge of results combined with other incentives. It is probably rare that a knowledge of progress operates alone. It is likely that such factors as rivalry and social recognition are always involved in some degree. But in the experiments so far reported, these other factors were not emphasized. In many experiments, however, the knowledge of progress has merely been taken as the *occasion* for utilizing other motives, such as praise and blame, rivalry, money or other rewards, and the like.

At least two studies have attempted to use a knowledge of intelligence scores as an occasion for verbal suggestion and other forms of motivation. Mitchell [47] divided the lowest fourth of the freshman class in a high school into two equivalent groups on the basis of the Otis tests. Each pupil in the experimental group received the following notice, without further comment:

> Dear Pupil:
>
> Your score on the Intelligence Test which was given at the opening of school is LOW. This will mean that much work and effort on your part will be necessary to keep up with the class. Put yourself to the task and show that you can do it. YOU CAN IF YOU WISH.
>
> Principal

At the end of the year it was found that 62 per cent of the group which had received this notice passed on all subjects, while only 15 per cent of the equally poor group, which had not been notified, did so.

---

[47] Claude Mitchell, "Why Do Pupils Fail?" *Junior-Senior High School Clearing House,* 9: 172–176, November, 1934.

The author [48] conducted a somewhat similar study with college students at the University of Kentucky. From the lowest fifth in intelligence, experimental and control groups of 40 freshmen each were formed upon the basis of psychological tests, sex, and fraternity affiliation. The students in the experimental group were then called together, and a frank statement was made regarding their scores. They were told that it was important at the outset to recognize the fact that they were up against a somewhat different situation from that of the students with higher test scores. They were assured, however, that the experience at the University showed that such students could succeed, if they were willing to work and did not attempt too heavy a load in school or too many activities outside. The control group had no advance information.

The record of the two groups is summarized in Table 32. The mean-point standing of the experimental group was .94 for the first semester and .85 for the second semester, while the corresponding values for the control group were .64 and .69, respectively. The difference between the two groups is five times its probable error the first semester and more than twice its probable error the second semester. During the first semester three times as many students in the experimental group as in the control group made a point standing of 1.00 or better, and more than twice as many made this standing the second semester. Approximately twice as many experimental as control students passed all subjects. On the whole, the difference was more marked for the first semester than for the second, and was decidedly greater for the College of Commerce than for the College of Arts and Sciences. These two studies offer rather convincing evidence that a tactful handling of intelligence test scores may have a wholesome motivating effect on low-ranking freshmen in high school and college. More recent studies have tended to confirm these findings.[49]

A great many more studies have utilized achievement test scores as occasions for various types of motivation. In a well-known study Book and Norvell [50] used a knowledge of results in four laboratory

---

[48] C. C. Ross, "Should Low-Ranking College Freshmen Be Told Their Scores on Intelligence Tests?" *School and Society,* 47: 678–680, May 21, 1938.

[49] *Cf.* R. K. Compton, "Student Evaluation of Knowing College Aptitude Test Score," *Journal of Educational Psychology,* 32: 656–664, December, 1941.

Edna E. Lampson, "How Objective Can Freshmen in College Be toward Objective Evidence of Their Ability and Achievement?", *Educational Administration and Supervision,* 28: 280–290, April, 1942.

[50] William F. Book and Lee Norvell, "The Will to Learn: An Experimental Study of Incentives in Learning," *Pedagogical Seminary,* 29: 305–362, December, 1922.

TABLE 32

POINT STANDING FOR THE FIRST AND SECOND SEMESTERS FOR LOW-RANKING FRESHMEN WHO WERE TOLD THEIR INTELLIGENCE TEST SCORES AS COMPARED WITH THOSE WHO WERE NOT

POINT STANDING	FIRST SEMESTER						SECOND SEMESTER					
	Arts & Sci.		*Commerce*		*Total*		*Arts & Sci.*		*Commerce*		*Total*	
	Exp.	Con.	Exp.	Con.	Exp.	Con.	Exp.	Con.	Exp.	Con.	Exp.	Con.
1.80–1.99....			1		1							
1.60–1.79....		1	1		1	1	2	3			2	3
1.40–1.59....	3	1			3	1	1	1	1		2	1
1.20–1.39....	4	2	1		5	2	2	2	3		5	2
1.00–1.19....	7	2	2		9	2	4		1		5	
.80– .99....	2	5	1	2	3	7	1	6	1	1	2	7
.60– .79....	4	5	3	3	7	8	3	3	5	1	8	4
.40– .59....	3	3	5	2	8	5	3	1	2	6	5	7
.20– .39....		3	2	7	2	10	1	4		3	1	7
.00– .19....	1	2		2	1	4	3	1	1	2	4	3
Total.......	24	24	16	16	40	40	20	21	14	13	34	34
Mean.......	.98	.78	.83	.45	.94	.64	.85	.88	.86	.44	.85	.69
S.D.........	.36	.41	.46	.25	.41	.39	.50	.49	.45	.22	.45	.46
M_E–M_C....	.20		.38		.30		—.03		.42		.16	
P.E. of Diff..					.06						.07	

experiments as a basis for building morale or developing the "will to learn." For example, students in the experimental groups "were frequently told that if they would only make up their minds to increase their score they would somehow find a way to do it," while at the same time the "method of measuring their output and having them keep track of their score usually convinced them that this was true." Their data support the conclusion that this "special group of incentives" helps the experimental group to "make more improvement with a given amount of practice than do the control groups." But it is impossible to tell just how important a knowledge of results by itself would have been.

An experiment by Hurlock,[51] which utilized test results as occasions for praise and reproof, has attracted considerable attention. The subjects were 106 children in fourth- and sixth-grade arithmetic. The groups were equated on an initial practice period of fifteen minutes. Four more practice periods were held on successive days. The control group received the tests without comment. The praised group had their names read aloud at the beginning of each practice period. They were then called to the front of the room and received praise combined with exhortation to do still better work. Then the names of the children in the reproved group were called, and they were severely reproved for poor work, carelessness, and general inferiority. The ignored group heard what was said to the others, but they received no recognition whatsoever. The results are shown in Figure 39. After the first day reproof seemed far less effective than praise, although somewhat better than being ignored altogether. The control group made no progress whatsoever. It is to be regretted that this excellently planned experiment was not continued for several days longer. Manifestly an hour's total working time is insufficient to establish fully the comparative merits of these incentives as they would operate day after day in the ordinary classroom.

---

[51] Elizabeth B. Hurlock, "An Evaluation of Certain Incentives Used in School Work," *Journal of Educational Psychology*, 16: 145–159, March, 1925.

In a somewhat similar experiment in the same grades, Hurlock [52] studied the effect of group rivalry on addition. The control group took their tests for ten minutes on four days without comment. The experimental group was divided into two equivalent subgroups which were pitted against each other. The author emphasized the fact every day that the two groups "were absolutely equal, and that one had as much chance to win as the other." Although the effect of rivalry was present in all types of pupils, it was most marked in younger pupils and in inferior pupils. Increase in accuracy was much less than increase in speed, with some tendency for increase in speed to be accompanied by reduction of accuracy. It is well to keep in mind Thorndike's warning that "the attainment of active rather than passive learning at the cost of practice in error may often be a bad bargain." [53] The study was too brief to be conclusive.

Another study [54] shows that repeated applications of praise or blame may have different effects on introverted and extroverted pupils. Introverted fifth-grade pupils improved faster in number cancellation exercises when praised than did either introverts who were blamed or extroverts who were praised. However, extroverted pupils when blamed improved faster than extroverts who were praised or introverts who were blamed. Unfortunately one

---

[52] Elizabeth B. Hurlock, "The Use of Group Rivalry as an Incentive," *Journal of Abnormal and Social Psychology,* 22: 278–290, October-December, 1927.

[53] Edward L. Thorndike and others, *op. cit.,* page 147.

[54] George C. Thompson and Clarence W. Hunnicutt, "The Effect of Repeated Praise or Blame on the Work Achievement of 'Introverts' and 'Extroverts'," *Journal of Educational Psychology,* 35: 257–266, May, 1944.

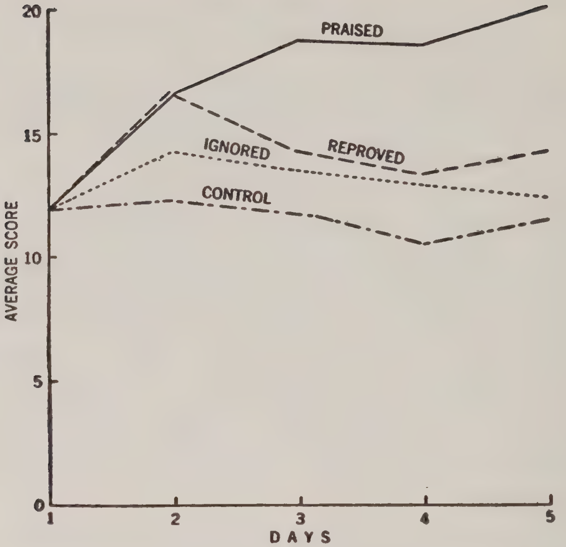


Figure 39. A Study of the Influence of Praise and Reproof upor Achievement in Fourth-Grade and Sixth-Grade Arithmetic. (Aftei Hurlock.)

cannot safely conclude from a study which involved a total practice time of three minutes upon highly artificial tasks that the same differences would necessarily appear under ordinary school conditions. The problem is worthy of further experimentation.

Two studies attempt to compare the relative effectiveness of individual and group rivalry. From an experiment with college students involving twelve two-minute practice periods in rate of substitution and twelve three-minute practice periods in improving reading rate, Sims[55] concluded that "individual motivation is vastly superior to group motivation and group motivation is but slightly superior to no motivation other than that which comes incidentally in learning." The group motivation involved two approximately equal groups competing with each other. It should be noted that these were artificial groupings rather than natural ones with unified morale or group consciousness. In the individual motivation the students worked in pairs, each student competing against another. Also, the names of the three students who showed the greatest improvement and the names of the three students who showed the least improvement and the three with the highest and the three with the lowest scores on the previous day's practice "were reported to the class with the praise or blame which the record deserved." Other experiments along this line need to be continued long enough for the novelty to wear off, as well as to test the permanency of retention. Apparently the novelty feature was more evident in the individual motivation group, as the other situation was much more like that of the ordinary classroom.

---

[55] Verner Martin Sims, "The Relative Influence of Two Types of Motivation on Improvement," *Journal of Educational Psychology*, 19: 480–484, October, 1928.

Maller [56] conducted a study of individual and group rivalry in which prizes were offered for the winners. The task was speed in solving arithmetic problems. The study included 1,538 children in grades five to eight in ten different schools. There were six two-minute practice periods for each type of situation. The speed of work for self was found to be "consistently and significantly higher" than that of work for the group, with the differences increasing as time went on. There were, however, marked individual differences. There was a low positive correlation between co-operativeness and mental age and between co-operativeness and chronological age. Maller found that co-operation with an organized team resulted in even greater efficiency than working for self. One of the most significant educational implications, both from Maller's study and from experiments in industry, is that the maximum of co-operation is likely to be obtained when the group is of highest homogeneity.[57] Zubin quotes experimental evidence [58] in support of the same point:

> Individual vs. group rivalry is affected by the individual's relative proficiency with respect to the rest of the group. Thus . . . an individual of known superiority will probably work much below his maximum capacity when competing with his inferiors in a given task.

The important point is that *co-operation can be and has to be learned.* Teamwork is itself a worth-while educational objective. There is good reason to believe that, with proper organization, measurement could motivate the group as hitherto it has stimulated the individual.

The aftereffects of specific connections. Most studies of motivation, such as those which have been summarized in the foregoing discussion, have been concerned with the learning of a series of related activities. Thorndike [59] and his students have also investigated the influence of the aftereffects of specific connections which they immediately follow and to which they belong. Thorndike feels that the results of numerous experiments conducted under his direction "prove that a satisfying after-effect of a connection can and generally does strengthen that connection directly, irrespective of repetitions or rehearsals or recalls of the connection and of images or other representations of the after-effect." [60] Annoying after-effects, on the other hand, do not necessarily weaken the connections to which they are attached, or may do so "by strengthening some competing connection."

---

[56] Julius Bernard Maller, *Coöperation and Competition,* 176 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1929.

[57] Goodwin Watson, "The Surprising Discovery of Morale," *Progressive Education,* 19: 33–41, January, 1942.

[58] Joseph Zubin, *Some Effects of Incentives,* page 4. New York: Bureau of Publications, Teachers College, Columbia University, 1932.

[59] Edward L. Thorndike and others, *The Fundamentals of Learning,* 638 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1932.

Rock [61] conducted experiments in code learning and ball tossing to determine the effect of variation in the amount of rewards and punishment. The study combined knowledge of results in specific connections, provided by the statement "Right" or "Wrong," plus rewards and penalties in terms of money. Although the punishment was distinctly less effective than were the rewards, "the results would indicate that this reaction is brought about almost as certainly, or with almost the same strength, by the least reward as by the greatest." As a matter of fact, the mere statement "Right" after a response appeared often to be as effective as when money rewards were added to it. In other words, fortunately for education, knowledge of success is itself one of the most powerful rewards. And, of course, there is always the possibility that the subject may be too highly motivated. More than one batter has fanned out at a critical point in the game because he was too anxious.

Forlano sought to determine the effect of monetary rewards both actual and promised, when combined with a knowledge of progress in learning spelling and arithmetic. He found that "the efficacy of the rewards was not as great as would be expected." In fact, the author makes this significant statement: [62]

> It is probable that conclusions concerning the operation of rewards may be true for one part of the field of learning but will not necessarily hold for other sections of this field. It may be that repetition or practice plus normal class discipline is all that is needed for efficient school learning.

A knowledge of results often has a definite *guidance* function which increases the amount of learning by influencing the direction of practice. Thorndike [63] presents convincing experimental evidence on this point. For example, 12 subjects practiced making

---

[60] *Ibid.,* page 270.

[61] Robert T. Rock, Jr., *The Influence upon Learning of the Quantitative Variation of After-Effects,* 78 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1935.

[62] George Forlano, *op. cit.,* pages 55–114.

[63] Edward L. Thorndike and others, *op. cit.,* pages 184–202.

3-inch, 4-inch, 5-inch, and 6-inch lines for seven training periods, in all 4,200 lines, when blindfolded and without a knowledge of results. The outcome was "approximately zero change." When six of these subjects were given further practice with knowledge of results, all showed improvement. In a later experiment Thorndike had 37 adults write with the hand they were not accustomed to use for 50 periods of five minutes each. Half the time the eyes were closed or bandaged, and the other half the eyes were open. With the eyes open, the subjects "showed a little less gain in speed and 2.66 times as much gain in quality."[64] Other experiments[65] have revealed a tendency for learning to increase with the increasing precision of the learner's knowledge of results.

In view of the close relationship ordinarily found between awareness of results and the progress of learning, one may raise the question of whether learning ever occurs without the awareness of what is being learned. Ordinary experience indicates that such learning does occur. Individual idiosyncrasies of posture and speech are examples. Thorndike and Rock find evidence of such learning under laboratory conditions, although Irwin and his associates question some of the assumptions involved.[66] At any rate, that such learning is at best risky and uneconomical is evidenced by the fact that about one-fourth of Thorndike's and Rock's subjects "probably did not learn at all."

II. The Relation of Measurement to the Type of Learning

Closely related to the *amount* and *quality* of learning is the *type* of learning, or the *learning procedure* which is employed. There is considerable evidence for thinking that effective work or study habits of the student are of fundamental importance in learning. A question of major importance, therefore, is: to what extent does the type of measurement used influence the type of study technique employed by the student? Some important studies bearing on this question have been conducted on the college level, but unfortunately none has so far been reported on the elementary or secondary school level.

In a pioneer study, Terry[67] found that 236 students in educational psychology were "influenced to a significant extent" by the type of examination for which they were preparing. The most striking characteristic of the methods employed in preparing for an objective test which had been announced a month in advance was the students' emphasis on details, while they tended to study for large units of subject matter when they were preparing for an essay examination announced for the next month. Douglass and Tallmadge [68] reported similar results at the University of Minnesota. They found that the "objective type focuses attention upon details and exact wording, while the subjective type apparently favors methods involving organization, perceiving relationships and trends, and personal reactions."

---

[64] Edward L. Thorndike and others, *The Psychology of Wants, Interests, and Attitudes,* page 125. New York: D. Appleton-Century Company, Inc., 1935.

[65] Margery Hayden Trowbridge and Hulsey Cason, "An Experimental Study of Thorndike's Theory of Learning," *Journal of Genetic Psychology,* 7: 245–258, October, 1932.

[66] See *Journal of Experimental Psychology,* 17: 1–19, 823–827, February and December, 1934.

[67] Paul W. Terry, "How Students Review for Objective and Essay Tests," *Elementary School Journal,* 33: 592–603, April, 1933.

There also appear to be significant differences among the various forms of the so-called new-type examinations in their effect on study methods. Terry [69] found, for example, that the one predominant method of preparing for completion tests emphasized the word-for-word mastery of statements considered important, while preparing for true-false tests involved methods which dealt primarily with definitions and detailed facts such as the authors and findings of experiments. The author's conclusion points out an important educational implication:

> The kind of test to be given, if the students know it in advance, determines in large measure both what and how they study. The behavior of students in this habitual way places greater powers in the teacher's hands than many realize. By the selection of suitable types of tests the teacher can cause large numbers of his students to study, to a considerable extent at least, in the ways he deems best for a given unit of subject-matter.

Meyer [70] conducted a careful laboratory experiment with 124 psychology students to determine the relation between the specific examination-set and immediate memory and delayed memory after five weeks. When the *amount* of study was held constant, the *method* and *results* appeared to be largely dependent on whether the set was for recall or for recognition tests. It appeared that when students expected completion tests they studied with more effort than they would have put forth for recognition tests. More students made summaries and maps and otherwise attempted to obtain a general picture of the material when they expected essay examinations than otherwise. Meyer points out four practical implications:

---

[68] Harl R. Douglass and Margaret Tallmadge, "How University Students Prepare for New Types of Examinations," *School and Society,* 39: 318–320, March 10, 1934.

[69] Paul W. Terry, "How Students Study for Three Types of Objective Tests," *Journal of Educational Research,* 27: 333–343, January, 1934.

[70] George Meyer, "An Experimental Study of the Old and New Types of Examination," *Journal of Educational Psychology,* 25: 641–661, December, 1934; and 26: 30–40, January, 1935.

> 1. Since it is more economical, when a given amount of time is spent in studying, to use a recall examination set for delayed recognition or immediate and delayed recall tests, recognition questions should be used in testing only when they form a part of the entire examination or when students are unaware that such questions are to be used exclusively.
> 2. If the teacher feels it necessary that the students be able to recognize certain materials for a short time only, then the indications are that a recognition examination set may be used. This means that the teacher must evaluate the material in his course very carefully, since recognition tests, if given indiscriminately, may have a deleterious effect on what the students ultimately retain of the course.
> 3. If the teacher feels it necessary that the students be able to recall isolated facts when specific cues are given as to the fact wanted, a completion examination set may be used with profit.
> 4. If the teacher wants the students to recall the material in an organized fashion and to know facts when cues are not given, the essay examination set should be used in preference to any objective type of examination set. Here again the teacher must evaluate the material which he presents in the light of what the student should learn from the course.

The following quotation from Monroe [71] suggests that the nature of the examinations emphasized by the teachers may influence the students' reactions much more than the objectives of the course:

> There has been much discussion of the importance of teachers formulating their objectives and, in response to the pressure of authority, they have spent many hours in formulating lists of immediate objectives, that is, the goals toward which students are expected to direct their efforts. Many of these lists merit commendation, but their influence upon students is practically nil in comparison with the influence of the tests administered. Students direct their efforts toward becoming able to respond to the tests they anticipate.

D. Some Educational Implications

Much of the experimental evidence on motivation has been fragmentary, some of it contradictory, and hardly any of it conclusive. But a few generalizations appear to have been fairly well established.

Implications for educational theory. In the first place, there is grave danger of premature and unwarranted generalizations in psychology and education. That it is hazardous to generalize from the laboratory experiment to the classroom application has been demonstrated in motivation experiments again and again. It is also dangerous to generalize from one age level to another. This is one of the greatest limitations of much of the experimental work on motivation. There is a great need for comparing the results of experiments made on the college level with results obtained from pupils on the elementary and secondary school levels.

---

[71] Walter S. Monroe, "Some Trends in Educational Measurement," *Twenty-Fourth Annual Conference on Educational Measurements,* page 32. Bulletin of the School of Education, Indiana University, Vol. XIII, No. 4. Bloomington, Indiana: Bureau of Coöperative Research, 1937.

In a concluding statement, Forlano sums up the situation as follows:[72]

> Finally, the experimenter who attempts to check the results of a laboratory experiment in a classroom situation has an important and responsible task; for it is when conclusions of laboratory studies are not tested in the classroom situation that the benefit or harm to education results. It becomes imperative, therefore, that there be a systematic schoolroom check of results as they issue from the laboratory.

In the second place, there are no fixed motivating categories such as knowledge of results, praise and blame, rewards and punishments, et cetera. Brenner states this point well:[73]

> The truth seems to be that there do not exist such psychological entities but that they do act in *specific situations, depending upon all the factors of the situation as a whole.* What in one situation may constitute praise, under certain other circumstances will be considered blame. The incentives derive their attributes, so to speak, from the situation in which they are active.

Implications for educational practice. Three points require brief mention. In the first place, the measurement program of the school influences both the teacher and the learner. It affects teaching emphasis and curriculum content, as well as the amount and quality of learning and the procedure employed. In the second place, no motivating factor operates universally. Both Chase and Hurlock, for example, found young children more susceptible than older children to the motivation used. In general, praise seems more effective upon the duller and socially inferior groups. Frequent testing also seems most helpful to weaker pupils. On the other hand, there is some evidence that blame and knowledge of results are more effective in the stronger groups. Even in similar age and social groups, however, marked individual differences appear as to the relative effectiveness of different types of motives, or even as to the effectiveness of the same motive used at different times. Brenner[74] warns against a

> stereotyped habit of motivation, for instance, always praising the children, always smiling and appearing pleased. This form of mechanized motivation is not adequate for increasing the performance of children, and it is doubtless harmful in its influence upon character building in children.

---

[72] George Forlano, *op. cit.*, page 101.

[73] Benjamin Brenner, *Effect of Immediate and Delayed Praise and Blame upon Learning and Recall,* pages 48–49. New York: Bureau of Publications, Teachers College, Columbia University, 1934.

[74] *Ibid.*, page 50.

In the third place, no motivating factor operates automatically. Test scores, at best, merely provide an occasion for praise or blame, reward or punishment, or some form of social recognition. The strategic place of the teacher is nowhere more in evidence than in motivation. In a fundamental sense, the role of the teacher is to stimulate and guide the learning process. Perhaps Brenner's concluding statement [75] does not put the matter too strongly:

> The facts about the usefulness of a motive in a certain learning situation will be furnished by educational psychology, but proper application of the incentive in a given situation depends upon the insight of the teacher. The effectiveness or worth of a teacher depends upon his ability to make adequate use of motivation.

SELECTED READINGS FOR FURTHER STUDY

Allen, Clinton M., *Some Effects Produced in an Individual by Knowledge of His Own Intellectual Level.* New York: Bureau of Publications, Teachers College, Columbia University, 1930. 98 pages.

Brenner, Benjamin, *Effect of Immediate and Delayed Praise and Blame upon Learning and Recall.* New York: Bureau of Publications, Teachers College, Columbia University, 1934. 52 pages.

Forlano, George, *School Learning with Various Methods of Practice and Rewards.* New York: Bureau of Publications, Teachers College, Columbia University, 1936. Part II.

Kirkpatrick, James Earl, "The Motivating Effect of a Specific Type of Testing Program," *University of Iowa Studies in Education,* 9: 41–68, June 15, 1934.

Maller, Julius Bernard, *Coöperation and Competition: An Experimental Study in Motivation.* New York: Bureau of Publications, Teachers College, Columbia University, 1929. 176 pages.

Rock, Robert T., Jr., *The Influence upon Learning of the Quantitative Variation of After-Effects.* New York: Bureau of Publications, Teachers College, Columbia University, 1935. 78 pages.

Thorndike, Edward L., and others, *The Fundamentals of Learning.* New York: Bureau of Publications, Teachers College, Columbia University, 1932. 638 pages.

————, *Human Nature and the Social Order.* New York: The Macmillan Company, 1940. Part I.

————, *The Psychology of Wants, Interests and Attitudes.* New York: D. Appleton-Century Company, 1935. 301 pages.

Young, Paul Thomas, *The Motivation of Behavior.* New York: John Wiley & Sons, Inc., 1936. 562 pages.

Zubin, Joseph, *Some Effects of Incentives: A Study of Individual Differences in Rivalry.* New York: Bureau of Publications, Teachers College, Columbia University, 1932. 60 pages.

---

[75] *Ibid.*, page 50.

CHAPTER XII

Practice

A. Some Important Principles of Learning

General characteristics of learning. The term *learning* always implies a modification of behavior brought about through experience or practice. From the standpoint of the learner three phases of the process may be differentiated. First, there is a problem, a situation for which the individual has no adequate ready-made response. With an animal it may be a strange maze or puzzle box. A pencil and paper would present a problem situation to a child who has not learned to write. In the second place, the learner usually reacts to the problem situation in a variety of ways. As a rule, after many trials that have not been successful, the learner hits upon the correct response. On succeeding trials the incorrect responses are eliminated and the successful response is developed. This nurturing and developing of the successful response is the third step in learning. Learning is always a personal matter; it is what the individual learner does for himself.

The principle of exercise or practice. It is a matter of common observation that practice is essential to the development and maintenance of skill. Every athlete and musician who attains distinction does so only after persistent practice and can maintain his skill only by "keeping in training." In fact, the familiar maxim, "Practice makes perfect," seems to be accepted by everybody except professional psychologists, who insist that the statement requires essential qualification. There is certainly no money-back guarantee with practice. At best it merely *tends to,* or provides the *occasions for,* improvement. Most adults, for example, will testify to the fact that in spite of years of practice in handwriting they have grown steadily worse all the time. Thorndike offers the cautious suggestion that "other things being equal, the oftener a situation connects with or evokes or leads to or is followed by a certain response, the stronger becomes the tendency for it to do so in the future."[1] He also emphasizes the fact that the mere repetition of a situation is wholly ineffective. "With the repetition of a *connection,* meaning thereby a situation and a given response to it, the case is different,"[2] however. But the connection must involve more than contiguity in time and space; there must also be the principle of "belongingness," or the awareness on the part of the learner that "this goes with that." Other writers have suggested still other qualifications, such as recency, intensity, congruity, and completeness. It is sufficient for our purpose to note the futility of blind practice. *Only correct practice is effective.*

---

[1] Edward L. Thorndike and others, *The Fundamentals of Learning,* page 6. New York: Bureau of Publications, Teachers College, Columbia University, 1932.

The principle of effect. It will be noted that Thorndike prefaced his so-called "law of exercise" with the elastic phrase "other things being equal." One reason for this qualification is his belief that the effect accompanying or following the exercise of a connection is even more important than the exercise itself. This principle holds that "what happens as an effect or consequence or accompaniment or close sequel to a situation-response, works back upon the connection to strengthen or weaken it."[3] Psychologists have generally been willing to accept the obvious influence of satisfaction in increasing the likelihood that a response will be repeated and hence the connection be strengthened through exercise, but many of them have found difficulty in understanding how it "works back" upon the connection itself. But many things are accepted by practical people that are not understood by the wisest theoreticians. The extensive experimental data of Thorndike and his co-workers, as well as the experience of teachers of all ages, point clearly to the fact that what follows practice or exercise, as well as what precedes it, is an important consideration in determining the effectiveness of practice, although its method of operation may be obscure. *Effective practice takes into account the effect of practice.*

Since practice is always preceded by some kind of motive and followed by some kind of effect, it is manifestly difficult, if not impossible, to determine in any given case which is more important. Indeed, when two or more factors are indispensable or inseparable, little profit can come from discussing their relative importance. The futile discussion of the relative influence of heredity and environment is a case in point. The important thing is their relationship. Such experiments as that of Symonds and Chase on "Practice vs. Motivation"[4] appear to set up an artificial antithesis. Even if their groups had been equated and other variables controlled and the experiment carried on for a longer period than twelve days, some doubt would still remain about the meaning, if not the validity, of their conclusion:[5]

---

[2] *Ibid.,* page 64.

[3] *Ibid.,* page 6.

[4] Percival M. Symonds and Doris Harter Chase, "Practice vs. Motivation," *Journal of Educational Psychology,* 20: 19–35. January, 1929.

> As to the practical implications we must conclude that the most effective device that can be applied to learning is to increase the amount of drill or practice. The prime function of motivation is to make this drill or practice more palatable.

After all, why must one choose practice *or* motivation, rather than practice *with* motivation? Even if it were not impossible to separate the two entirely, their co-operation is most desirable educationally.

Relation of measurement to practice in learning. It would appear, therefore, that measurement is related to practice in two ways. The awareness that the tests are to be given may stimulate study or drill on the material to be covered. This possible motivating effect has been considered in the preceding chapter. The test also provides an opportunity for learning by practice. The present chapter will consider the educational value of the actual taking of the test or examination. To what extent are examinations of all types like clinical thermometers which merely measure the present status of the patient and which, in and of themselves, are wholly incapable of raising or lowering the temperature? It is convenient to treat separately the practice effect of pre-tests, intermediate or class tests, and final examinations.

B. The Educational Value of Pre-Tests

Why pre-tests? There are two primary reasons for giving pre-tests. In the first place, they are commonly used in educational research to determine the status of the experimental and control groups at the beginning of the experiment. The results of these tests are used to determine the equivalence of the groups and, by comparison with similar tests used at the end, to determine the amount of change which has occurred. In the second place, pre-tests are used in educational guidance to reveal the strong and weak points of the pupils at the beginning of a period of instruction, and hence to serve as the basis for remedial teaching. The order of events for each instructional unit according to the so-called Morrison mastery formula is: "Pre-test, teach, test the result, adapt procedure, teach and test again to the point of actual learning."[6] It is, of course, clear that pre-tests expose the pupil to many items which are new to him. What evidence is there regarding the learning which results from these preliminary tests? It would appear that there are possibilities both for negative and for positive learning: that is, the tests may be a help or a hindrance.

---

[5] *Ibid.*, page 34.

[6] Henry C. Morrison, *The Practice of Teaching in the Secondary School* (Revised Edition), page 81. Chicago: The University of Chicago Press, 1936.

Danger of negative learning. Many writers have expressed serious fear as to the danger of negative learning in pre-tests, especially if the tests are of the true-false type. For this hypothesis, however, there has been more argument than evidence. These arguments are largely based on the factor of primacy in learning, on the belief that first impressions are lasting, and on the evil effects of practice in error. As approximately half of the statements in a true-false test are false, there would appear to be real danger to the pupil from erroneous first impressions. But there is also good psychology on the other side. Modern psychology recognizes the importance of the total situation or configuration in learning. Whether or not a false statement is dangerous depends largely upon the setting in which it appears. A false statement in the textbook, toward which the characteristic pupil attitude is likely to be one of passive, uncritical acceptance, might easily be serious. But the situation is different with the items in a true-false test. Here the habitual attitude of the modern pupil is one of active, critical challenge. "Practice fixates or disrupts acts," says Carr, "according to the circumstances under which they are performed."[7] Dunlap has shown that the determining factors are the thoughts and desires of the learner, which, when properly directed, may actually establish the *response directly opposite to the one practiced.*[8] Whether a response, correct or incorrect, is helped or hindered by practice depends upon the total situation; in this the attitude of the learner is most important.

In the final analysis, however, the question must be settled by actual experimental evidence rather than by an appeal to psychological theory. In 1929 Ruch summarized this evidence, which, though "insufficient in quantity," appeared to warrant two important but "highly tentative" conclusions:[9]

> 1. The negative suggestion effect of false statements in true-false tests is probably much smaller than is sometimes assumed.
> 2. The small amount of negative suggestion which has thus far been shown for true-false tests seems to be fully offset by the net positive teaching effects.
> 3. Such harmful effects as sometimes occur can be, for the most part, if not entirely, prevented by having the students correct their papers in class.

---

[7] Harvey Carr, "Teaching and Learning," *Journal of Genetic Psychology,* 37: 210, June, 1930.

[8] Knight Dunlap, *Habits, Their Making and Unmaking,* 326 pages. New York: Liveright Publishing Corporation, 1932.

[9] G. M. Ruch, *The Objective or New-Type Examination,* page 368. Chicago: Scott, Foresman & Company, 1929.

The experimental evidence published since that time has tended to confirm these conclusions and to add one other important conclusion: [10]

McIntosh [11] reports a study of a somewhat related problem: What is the effect upon learning of introducing the incorrect form into the practice situation? Two school subjects, grammar and spelling, were used with junior-high-school material in this six-week experiment. Both experimental and control groups studied the same lesson sheets for a prescribed time. Half of the groups were then given mimeographed exercises of the modified true-false type with instructions to correct any errors found. Comparable groups were given ordinary completion exercises covering the same content. No statistically significant differences were found in the test results. The author concludes that under the condition of the experiment "exposure to wrong forms results in no observed disadvantage to the learners." Furthermore, McIntosh found practice exercises in spelling, in which the pupils merely recognized the correct forms, were definitely inferior both to exercises requiring pupils to correct any spelling errors found and to exercises requiring the recall of the correct forms.

Evidence of positive learning. Unfortunately, relatively little attention has been given to determining the positive value of pre-tests, aside from the effort to show that the negative effects of true-false tests are fully offset by the positive effects. But there are many forms of objective tests, and only the true-false has been seriously suspected of inducing negative learning. Reference has already been made to a study of the use of pre-tests to introduce each unit in high-school physics, as reported by Kirkpatrick.[12] The tests were corrected in class, and were used as a basis for class discussion and subsequent study. The results were most beneficial to the pupils in the lowest third in mental ability, although there was a statistically significant difference in favor of the experimental group as a whole, when measured by tests of physics information and comprehension.

Jersild [13] has made a study of the value of pre-tests in teaching beginning psychology. He found the pre-tests of the true-false type used to introduce a regular unit of class work and outside reading showed "partly negative and essentially unreliable advantages." However, when a pre-test of the multiple-choice type was compared with the same items in oral and written summaries, and a pre-test of the brief essay type was compared with a preliminary topical outline, he found a "consistently higher score for the pre-examined group." The advantage varied from 5 to 20 per cent. The author's conclusion was that this result was due to the fact that this "direct interrogation" had the effect of "stimulating the industry of the learner." Insofar as this is a correct interpretation, the value of the pre-test consisted more in motivation than in practice or drill. Since the selections used in the experiment were read by the instructor to the class and were followed *immediately* by the end-tests without an opportunity for study, it is not very clear how the "industry of the learner" had a chance to show itself.

---

[10] See *Journal of Educational Psychology,* 25: 281–285; 422–426, April and September, 1934; and *Journal of Experimental Education,* 2: 269–273, 1934.

[11] John Ranton McIntosh, *Learning by Exposure to Wrong Forms in Grammar and Spelling,* 61 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1944.

[12] James Earl Kirkpatrick, "The Motivating Effect of a Specific Type of Testing Program," *University of Iowa Studies in Education,* 9: 41–68, June 15, 1934.

[13] Arthur T. Jersild, "Examination as an Aid to Learning," *Journal of Educational Psychology,* 20: 602–609, November, 1929.

Keys has also produced some experimental evidence in support of the proposition that subjecting students to erroneous statements helps to produce a "generally critical attitude toward unproved assertions of a psychological character."[14] Further evidence on this point is badly needed, for one of the greatest services any educational agency could offer would be to render future citizens immune to the endless propaganda of politicians, advertisers, and salesmen, in which our restless world abounds.

C. The Educational Value of Intermediate or Class Tests

Why class tests? A distinction should be made between two functions of tests and examinations. The first function is that of *measurement.* Good illustrations are college entrance examinations and final examinations for promotion and graduation. The second function is that of *instruction.* Tests given by the teacher primarily for educational diagnosis and remedial instruction or for drill are illustrations of the second type. Ordinary class tests may fall in either classification. Many teachers feel that their chief dependence for determining the mark in the course must be on written examinations and that a fairer sampling of the pupil's performance is to be had from tests distributed throughout the semester than can be had from the final examination alone. Other teachers use the class tests to motivate regular study and review, to guide teaching, and to provide a specific type of educational experience which comes from taking them. Both functions are important. While it is quite unlikely that the same test will be equally good for both purposes, it is usually true that a test designed primarily for the purpose of measurement will have instructional value also, if the results are properly handled. Conversely, the results of tests designed primarily for instructional purposes can generally be utilized to some extent for determining the pupil's record in the course. A test has a legitimate place, however, if it can be shown that it does either job well. A considerable body of experimental literature gives ample testimony to the possibilities of tests as tools of instruction.

---

[14] Noel Keys, "The Influence of True-False Items on Specific Learning," *Journal of Educational Psychology,* 25: 511–520, October, 1934.

Tests in remedial instruction. Tests have a threefold relationship to remedial instruction. In the first place, they are used to reveal the need for remedial instruction. Tests are important tools in educational diagnosis.[15] In the second place, many instructional materials are in the form of exercises that are not unlike tests. Diagnosis without remedial treatment is as ineffective in education as in medicine. In the third place, tests are used to determine when the need for remedial instruction has been met.

Smith[16] found that the use of practice tests enabled children in beginning reading to read as well at the end of the first semester as average children do at the end of a year. Maloney and Ruch[17] compared the relative effectiveness for a period of ten weeks of the ordinary textbook method of instruction in grammar, ten tests of 25 items each with no textbook, and a combination of the textbook method with four or five tests. The subjects were 497 pupils in the ninth, tenth, and eleventh grades. Although the pupils spent about 40 per cent less time in study by the test method of instruction, it proved on the final examination, to be the best of the three methods. The straight textbook method was the poorest of all. Whether the superiority of the test method was due to its novelty or to the greater definiteness of the assignments, the authors were unable to say. A study[18] on the college level indicates that students may become too dependent on such study aids. Weekly objective tests used as study aids resulted in a superiority of from 12 to 15 per cent on the instructor's unit tests over a class taught without such aids. The advantage had entirely disappeared, however, on the final examination, when the students were not allowed access to the study tests for review. Symonds[19] compared the effectiveness of grammatical instruction of the textbook type with drill exercises similar in form to test items, and with a combination of the two. The subjects were three sixth-grade classes, and the period of instruction was 15 minutes per day for three weeks. His conclusion has wide implications:

---

[15] This function of tests will be considered more fully in Chapter XIII.

[16] Nila Banton Smith, "An Experiment to Determine the Effectiveness of Practice Tests in Teaching Beginning Reading," *Journal of Educational Research,* 7: 213–228, March, 1923.

[17] Estelle L. Maloney and G. M. Ruch, "The Use of Objective Tests in Teaching as Illustrated by Grammar," *School Review,* 37: 62–66, January, 1929.

[18] O. E. Hertzberg, J. D. Heilman, and H. W. Leuenberger, "The Value of Objective Tests as Teaching Devices in Educational Psychology Classes," *Journal of Educational Psychology,* 23: 371–380, May, 1932.

> One thing stands out clearly: It is the quality of drill or practice that counts and not its amount. Mere mechanical repetition apparently yields almost no learning. Instead of increasing the amount of drill in school subjects, more attention should be paid to the nature of the drill.

This is an important generalization well supported both by psychological theory and by experimental evidence. To assert that the *quality of practice* is more important than its *amount* is not to deny that the latter is also important. It does mean, however, that first consideration should be given to the quality of practice. What, then, are the characteristics of practice materials that are effective in learning? Symonds suggests one quality by implication. Repetition should not be mechanical; on the contrary, it should be animated and lively. This fact suggests that after all it may be the quantity of motivation behind the practice that makes the practice significant.

Studies in the same field by Cutright,[20] however, show that *not* all forms of motivated drill are equally effective. In one study of sixth-grade pupils in 17 schools, she found that drill games were relatively ineffective in correcting 50 common errors in English usage. A dramatized attack which consisted of writing plays, making slogans, or preparing programs was very much better, but the best procedure of all was that in which the pupils were furnished lists of their own individual errors, for the correction of which they planned their own remedial projects with the aid of the teachers. In a somewhat similar study involving over 1,800 pupils in the fourth, fifth, and sixth grades, she found that the use of drill games was the poorest of six methods tried, all involving the same amount of time. The most successful teaching method consisted of writing the selected form of grammatical usage in the blanks in specially prepared materials for drill, plus the oral reading of all sentences when completed. Miss Cutright's investigations demonstrate the soundness of two teaching principles of wide applicability:

1. That practice is most effective which is most directly related to the specific needs of individual pupils.

---

[19] Percival M. Symonds, "Practice Versus Grammar in the Learning of Correct English Usage," *Journal of Educational Psychology,* 22: 81–95, February, 1931.

[20] Prudence Cutright, "A Comparison of Methods of Securing Correct Language Usage," *Elementary School Journal,* 34: 681–690, May, 1934.

2. That practice is most effective in which the responses made are most nearly like those called for in actual life.

Recitation versus rereading in study. Pupils should be taught the value of self-imposed tests. One of the best ways for the pupil to prepare for an examination is to prepare one of his own of the type he expects the instructor to give. It is still better to exchange his test with those similarly prepared by other pupils and to discuss the results. Many pupils never learn the value of the recitation method of study. Gates [21] found, for example, that half or more of the time devoted to study of short selections may profitably be employed by the elementary school pupil in reciting to himself and in checking his responses with the copy.

A study by Forlano [22] involving 623 pupils in the fifth and sixth grades "working under ordinary schoolroom conditions and with methods and materials common to the school" came to the conclusion that "learning by recitation is clearly superior to learning by reading." Another study by Peterson [23] on the college level verified the value of recall but indicated that the amount of time required to read the selection must be taken into account. He concluded that "in the ordinary study situation a smaller proportion devoted to recall than that indicated by the Gates experiment will usually be found advisable."

A study by Spitzer [24] involving 3605 sixth-grade pupils in nine Iowa cities justified the conclusion that objective tests are effective learning devices which definitely aid retention. A later study by Sones and Stroud [25] which included 1300 seventh-grade pupils in Iowa showed rather clearly that reviews in the form of objective tests are most effective when used immediately after the original learning and decrease in value as time passes. Another study by Stroud and Freeburne [26] established the fact that an objective test response functions as a review medium in the same manner as direct recall.

---

[21] Arthur I. Gates, *Psychology for Students of Education* (Revised Edition), pages 335–337. New York: The Macmillan Company, 1930.

[22] George Forlano, *School Learning with Various Methods of Practice and Rewards*, pages 9–54. New York: Bureau of Publications, Teachers College, Columbia University, 1936.

[23] H. A. Peterson, "Recitation or Recall as a Factor in the Learning of Long Prose Selections," *Journal of Educational Psychology*, 35: 220–228, April, 1944.

[24] Hubert F. Spitzer, "Studies in Retention," *Journal of Educational Psychology*, 30: 641–656, December, 1939.

[25] A. M. Sones and J. B. Stroud, "Review, with Special Reference to Temporal Position," *Journal of Educational Psychology*, 31: 665–676, December, 1940.

[26] J. B. Stroud and Max Freeburne, "Symbolic Practice," *Journal of Educational Psychology*, 33: 65–71, January, 1942.

Devices for increasing instructional value of class tests. It is commonplace to say that learning is an *active* process and that pupils learn from *their own* activity rather than from that of the teacher. Yet this principle is often violated in testing as well as in other aspects of the teaching process. For four successive years Curtis and Woods[27] made a study of the relative teaching value of four common practices in correcting examination papers in various science classes in the junior and senior high school. The methods compared were as follows:

1. Pupils corrected own papers while the teacher read the correct answers. Free discussion followed.
2. Teacher checked incorrect items but made no corrections. Papers later returned and discussed item by item.
3. Teacher carefully wrote in all corrections. Papers later returned and discussed item by item.
4. Teacher carefully wrote in all corrections. Papers later returned, but only the questions pupils asked about were discussed.

The tests, which consisted of 100 items representing various objective forms, were repeated without warning the next day and after six weeks. The results indicate that the fourth method was the poorest, the second and third about equal in value, and the first method was the best of all. It should be noted that in the best method the teacher is least active and the pupils most active, and that in the poorest method the conditions are reversed. And yet, this obvious fact, that improvement results fundamentally from the pupil's own efforts, is one of the most difficult things teachers have to learn.

Curtis and Woods found that the errors in scoring varied from 5.7 per cent in the seventh grade to .4 per cent in the twelfth grade. There of course remains the problem of cheating. This would probably not be serious if the pupils were assured that the tests were to be utilized exclusively for diagnosis and drill and as such had nothing whatsoever to do with their marks in the course. With a reduced emphasis on school marks, many tests could be so used. The only way the test results could then affect a pupil's final mark would be indirectly, by providing him with the practice necessary for increasing his knowledge and skill. The only person who would be cheated, under such circumstances, by looking at another's paper or by incorrectly scoring his own paper, would be the pupil himself. It would then be just as foolish to cheat as to employ someone else to do one's tennis practice with the hope of improving one's own game.

---

[27] Francis D. Curtis and Gerald G. Woods, "A Study of the Relative Teaching Values of Four Common Practices in Correcting Examination Papers," *School Review,* 37: 615–623, October, 1929.

Teachers frequently wish to utilize the results of class tests both for instruction and for measurement purposes. In such cases the temptations to cheat are undoubtedly greater. How can the problem of cheating be solved when pupils are allowed to score their own papers? The solution usually proposed involves the use of some mechanical device or of separate answer sheets.

Use of test and drill machines and devices. Little[28] compared the ordinary method of teaching educational psychology, in which the class tests given at the end of each unit were scored by the instructor and were returned the next day, with two other methods which provided a knowledge of results immediately. One of these methods employed Pressey's device for the automatic scoring of objective tests and the tabulation of results item by item. As soon as the last student's paper came in, the distribution of scores was placed on the board and a discussion followed of the items missed by any considerable number. The other method employed Pressey's device for drill, which records the number of trials required by each student to make an errorless paper. All students in the two experimental sections who scored below *B* were required to take a make-up test conducted in the same manner the next day. The scores on the two tests were averaged. The results for the three methods are presented in Figure 40. The experimental groups excelled the control group throughout the distribution, although the difference was most marked in the lower half, where the students had the additional practice afforded by the make-up tests. In general, there was a slight advantage in favor of the drill-machine method over the test-machine method. It is possible that this advantage might have been greater had the class tests been of the multiple-choice type instead of the true-false type, where drill was at the minimum.

It should be stated that Little's study made no special effort to prevent cheating, and there was evidence that it did occur in some degree. The drill machines, however, did rather effectively control the factor of cheating owing to the students' scoring their own papers. But there still remained the possibility that their neighbors' behavior may have influenced the original response. All that would be necessary to reduce or to eliminate altogether that source of cheating is the preparation of different *arrangements* of the test items, and to take care that they are distributed in such a way as to prevent students seated together from receiving the same form of the test. Item 1 on one form, for example, might be item 25 on

---

[28] James Kenneth Little, "Results of Use of Machines for Testing and for Drill, Upon Learning in Educational Psychology," *Journal of Experimental Education,* 3: 45–49, September, 1934.

EXAM SCORE	TEST MACHINE f	DRILL MACHINE f	CONTROL f
185	1	1	
180	1	3	2
175	1	2	3
170	8	4	10
165	8	6	10
160	11	7	8
155	9	9	14
150	10	8	10
145	10	10	14
140	12	6	12
135	6	3	11
130	6	3	19
125	9	2	17
120	4	3	11
115	1	3	7
110	1	1	8
105			8
100			5
95			1
90			
85			

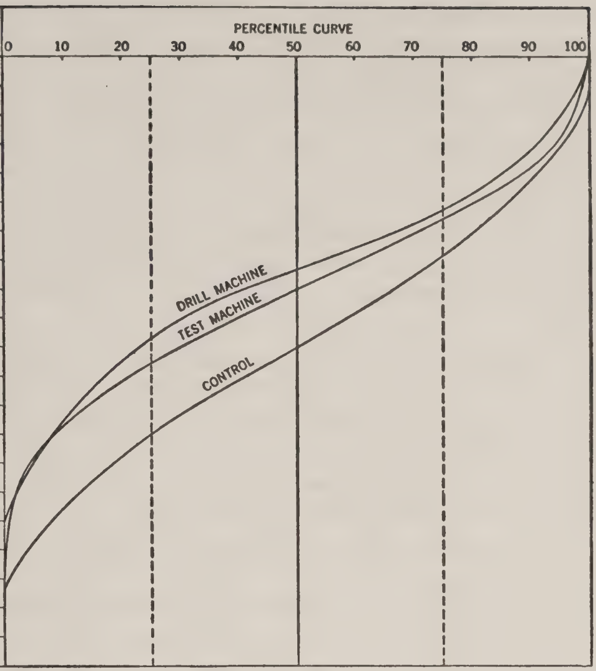


Figure 40. Test-Machine and Drill-Machine Methods of Teaching Educational Psychology Compared with Ordinary Methods. (After Little.)

another form, and so on, for such tests as true-false and recall. The items in multiple-choice and matching tests may be in the same order on all papers, but the answers should appear in a different sequence. In larger classes particularly, these different arrangements of test items serve a useful purpose. Each arrangement requires a separate scoring key, but the actual labor of scoring the papers is not materially increased. When the teacher reads the answers and the pupils score their own papers, however, the difficulties are greater.

J. C. and H. J. Peterson have devised two interesting methods of combining testing with self-instruction. Both are designed for recognition tests of the true-false and multiple-choice types. One consists of a specially prepared answer sheet for 100 items, with spaces for four or five possible answers to each. The sheet is printed with a kind of ink which turns one color when the correct answer is selected and a contrasting color when incorrect answers are selected. The student indicates his answers by touching them with a moist pencil or other instrument. As he is instructed to keep on marking until he gets the correct answers, the number of responses required is clearly indicated. The second method utilizes an answer form which consists of two layers of paper with a stiff cardboard between, with holes in it to correspond to correct answers. The student records his answer by punching the top sheet with a stylus. If his choice of answer is correct, the stylus goes through and punches the second sheet also. If his choice is incorrect, the stylus stops when it goes through the top sheet. The number of punches required to hit all the correct answers is, therefore, indicated on the top sheet. Both of the Peterson devices serve the functions of testing and of learning. The number of items correctly marked on the first trial, and so having only one response each, is a test record in the usual sense, while the total number of trials required is a record of learning. The authors [29] report several experiments with college classes which show statistically significant differences in favor of the self-instructor method of study. It has also been found that students prefer this method and are stimulated to ask the instructor many more questions than when the ordinary methods are employed.

Other devices for preventing cheating. Smeltzer [30] has also suggested a plan for having pupils mark their own papers in class that eliminates the danger of cheating. Each pupil is given a test paper with an answer sheet numbered to correspond to the number on the test paper. The pupil then records his answers on the answer sheet but does not write his name on it. The teacher next collects the answer sheets by rows and redistributes them in a different order. The teacher then reads the answers while the pupils mark the papers. A distribution of scores is put on the board from a show of hands by the pupils to indicate the number of pupils making the various scores. The number of pupils missing each item is indicated in a similar manner. Items causing difficulty are discussed. Papers are then returned to original rows and owners. Each pupil finally records his name on the answer sheet just before turning it in to the instructor. This scheme prevents cheating and presents an immediate knowledge of results for class discussion, but has the disadvantage that each pupil marks another pupil's paper rather than his own.

With the aid of student teachers in high school chemistry covering a period of three years, Hoff [31] found that pupils checking their own true-false class tests are "too inaccurate for just grading," since the percentage of error varied from 6.94 to 8.10. Cheating was a problem even with the brighter and more capable pupils, although to a smaller degree. The percentage of error was reduced to .61 (which was a better record than that made by the student teachers) when the following precautions were observed: true-false items were marked + and 0, rather than T and F; multiple-choice items were underlined as well as numbered; omitted items were crossed out before checking began; pupils exchanged papers with persons not personal friends; and the checker signed his name in a designated space.

---

[29] See *Transactions of Kansas Academy of Science,* 33: 41–47, 1930; 34: 291–296, 1931; 35: 132–140, 1932; also *The Kansas Industrialist,* June 14, 1933.

[30] C. H. Smeltzer, "Educational Engineering in Testing and Diagnosis," *Educational Method,* 12: 526–530, June, 1933.

[31] Arthur G. Hoff, "A Study of Honesty and Accuracy Found in Pupil Checking of Examination Papers," *Journal of Educational Research,* 34: 127–129. October, 1940.

The author has found that a variation of the foregoing systems makes it possible to have the pupil score his own paper. All that is required is to have the pupil fill out an answer sheet and turn it in to the instructor before the papers are marked. This gives a check on the correctness of scoring. If an ordinary pin is used to punch a hole to indicate the answer on recognition tests, a plain sheet of paper attached to the test or answer sheet will make available a duplicate copy of the pattern of answers that can be turned in immediately. Any change made on the paper after the duplicate has been turned in can be quickly detected. Answers to recall questions will have to be copied on the second sheet, unless a carbon sheet has been inserted between the two sheets. With this system there is very little likelihood of cheating when the pupils mark their papers as the teacher reads the answers.

Varying the form of the answer sheets and mixing the order when distributed will make it possible to dictate true-false and multiple-choice tests to pupils with small possibility of their copying from their neighbors. The plan works equally well with tests in mimeographed form. Usually two or three variations in mimeographed answer sheets will suffice. One form of answer sheet may direct the pupil to write 1 for true, and 2 for false. Other forms of the answer sheet may suggest 2 for true and 1 for false, 3 for true and 4 for false, and so on. Similar variations are indicated for multiple-choice items. It is also possible to have each pupil punch his answer with a pin upon duplicate answer sheets, a copy of which is then turned in before the original is scored by the pupil in class.

It will, of course, often be unnecessary to take these precautions to prevent cheating, especially when the *instructional* functions of the tests are stressed. The ideal prevention of cheating is a high morale among the pupils of the school. The above suggestions are made, however, to indicate that the possibilities for securing the instructional values which accrue when the pupils mark their own papers can be had even when cheating is a problem in the school.

Practice materials commercially available. For years commercial publishers have attempted to provide practice materials in the tool subjects. Among the earliest and best known were the Courtis[32] practice tests in arithmetic and handwriting and the Economy Remedial Exercise Cards[33] in arithmetic. Later came a veritable deluge of workbooks in many school subjects. These workbooks sought to help the pupil to help himself. Unfortunately, most of these publications took little account of the enormous extent of individual differences, or made adequate provision for motivation.[34]

In recent years some publishers have attempted to correct these weaknesses. A good example is the Strathmore Plan[35] which provides a closely integrated program of diagnostic tests and practice materials in the fundamental skills of arithmetic and English based upon the formula: "TEST—TEACH—PRACTICE—TEST." The underlying philosophy is clearly stated by the company's educational adviser, Dean Frank N. Freeman, as follows:

> The child first goes through inventory tests which reveal the essential items he does not know. He then studies the teaching exercises and practices on the exercises which are keyed to the parts he needs to learn. The teaching and his practice are quite specific and he is directed to the filling in of gaps which he himself has discovered in his own knowledge and skill. The connection between this knowledge and the successful performance of practical activities is brought out by the teacher and by the class discussion, and this gives meaning to the teaching and practice. The interest which is stimulated by this understanding of the meaning of the practice and by the pleasure in mastery may be enhanced, if necessary, by a judicious use of competition. Thus, this plan accomplishes a socially desirable end by a psychologically and educationally suitable means.[36]

Convenient lists of practice materials commercially available are to be found in recent books on diagnostic and remedial teaching, of which Blair[37] and Traxler[38] are good examples. Teachers gifted with imagination and industry can prepare similar practice materials of their own. One ingenious teacher[39] has made effective use of the daily newspaper in remedial English classes.

---

[32] Published by World Book Company, Yonkers, New York.

[33] Published by Scott, Foresman & Company.

[34] *Cf.* Earl P. Andreen, "A Study of Workbooks in Arithmetic," *Journal of Educational Research,* 32: 108–122, October, 1938.

[35] Published by the Strathmore Company, Aurora, Illinois.

[36] *Presenting the Strathmore Plan,* page 13. Aurora, Illinois: Strathmore Company, 1941.

[37] Glenn Myers Blair, *Diagnostic and Remedial Teaching in Secondary Schools,* 422 pages. New York: The Macmillan Company, 1946.

[38] Arthur E. Traxler, *The Use of Test Results in Diagnosis and Instruction in the Tool Subjects,* 80 pages. New York: Educational Records Bureau, 1942.

[39] Jesse Stuart, "Beginning and Eternal Ending: Making the Sparks Fly in a Remedial English Class," *Journal of the National Education Association,* 29: 131–132. May, 1940.

D. Educational Value of Final Examinations

Why final examinations? The attitudes of teachers vary widely regarding the educational function of final examinations as distinct from their administrative function of determining promotion, honors, and the like. Attitudes of teachers range all the way from regarding all examinations as relics of barbarism, inflicted upon innocent pupils as a form of retribution, to considering the taking and passing of examinations as the ultimate aim of all academic endeavor. In between are various shades of opinion and surprisingly little genuine evidence.

The genial but anonymous author of *The Psychology of Getting Grades* [40] stresses the practical aspects of the technique of grade-getting, in which the ability to "pass" some final examination is regarded by many instructors as *the* important thing. This author argues that "just as your grade depends on your ability to please the prof, success on the job depends on your ability to please the boss." He continues:

> The man who can sell his abilities to his superiors gets the grades and the diploma; he gets the job, he gets the raise; he gets the boss's daughter, the political office, and the big funeral.[41]

It is probable that the written examination affords fewer opportunities for the needed training than the oral, and that the least fruitful field of all is the objective test. Indeed, that is the unique merit of the objective test; it seeks to measure academic achievement and nothing else. The development of desirable traits of personality is a legitimate educational objective, but their measurement requires specially designed instruments.

Professor McClusky [42] is impressed with the possibilities of practice in error and of false impressions that may be left in the learner's mind by the ordinary final examination. He suggests that the solution is not to abandon the so-called final examination but to have it come earlier. McClusky recommends that

> the final examination should be conducted at least one if not two or more periods before the conclusion of a course in order to provide an opportunity to discuss the content of the examination. In other words, examinations should never be *final* —they should be *next-to-final*. Only carefully planned and stimulating discussions in a final effort to integrate the basic outcomes of the course should be final.

---

[40] Published by Lucas Brothers, Columbia, Missouri, 1935.

[41] *Ibid.*, page 89.

[42] Howard Yale McClusky, "An Experimental Comparison of Two Methods of Correcting the Outcome of an Examination," *School and Society*, 40: 566–568, October 27, 1934.

Nor are school administrators by any means agreed as to the educational value of final examinations. Ex-President Lowell of Harvard,[43] for example, holds that "one matter that has not yet received the attention it deserves is that of examinations as a vital factor in the educational process." He argues that final examinations are educationally valuable in these "three notable ways: by setting a standard; by requiring the expression of thought; and by promoting the association of ideas." It is obvious that at least two of these alleged values refer to the essay examination. Dean McConn [44] of New York University, argues that we have "professionalized" the final examination and that the remedy is to restore its lost "amateur status." Strictly speaking, he would substitute for the usual single final examination a multiplicity of "examinations and tests of many kinds, using them frequently, *but always informally, casually, and skeptically,"* and with the emphasis on instruction rather than on measurement. The emphasis placed by these administrative officials upon the instructional value of examinations is significant.

Unfortunately the advocates of final examinations, particularly the champions of essay examinations, have relied too much upon argument and not enough upon experiment. The few studies so far reported have investigated the value of objective tests. Scott,[45] for example, conducted an extensive study involving 805 junior and senior high-school pupils. He found in 34 out of 37 classes a "completely reliable difference in favor of using both standard tests and teacher-made tests as aids to learning," and concluded that "the use of final tests is worth while and that the time consumed by the teachers and students in the preparation and taking of such examinations is well spent." Other studies of the problem are greatly needed.

Some general considerations. To obtain the maximum educational value from final examinations, certain basic principles should be observed:

1. The educational value of all examinations, including the so-called final, should be stressed. This means that the examination should come early enough to permit sufficient time afterward for class discussion to clear up erroneous impressions. No exemptions should be permitted, and pupils should have training in preparing for and in taking tests and examinations.

---

[43] A. Lawrence Lowell, "Examination in the Educational Process," *Harvard Teachers Record,* 3: 184–188, October, 1933.

[44] Max McConn, "Measurement in Educational Experimentation," *Educational Record,* 15: 106–119, January, 1934.

[45] Ira O. Scott, *Stimulating Learning Through the Use of the Final Examination.* Unpublished Doctor's Field Study, 1937. Greeley, Colorado: Colorado State College of Education.

2. The major difference between the final examination and the regular class tests is in length. It is merely the last of a series of tests whose general purpose is the same. Increasing the frequency of testing, with the consequent reduction in the value of any one test, will go a long way toward eliminating the terror which the "final," under the present system, arouses in many students. The remark of the old woman about other women who comb their hair every day is appropriate here. "I just don't see how they stand it," she said. "I comb mine once a week and it nearly kills me."

3. The conditions under which the examination is held are important. An effort should be made to keep the conditions normal, as much like ordinary tests as possible. Whenever essay questions are used, sufficient time should be allowed for pupils to think carefully through and to organize their discussions before the actual writing, and to proofread them afterward. This in itself will do much to improve the quality of the examinations as an exercise in English composition. It must be kept in mind always that only correct practice is helpful.

The experimental evidence summarized in this chapter supports the generalization offered by Lindquist: [46] "The influence upon instruction of any specific testing program in any school situation primarily depends, then, upon the nature and quality of the examinations provided and upon the intelligence with which these examinations are used."

SELECTED REFERENCES FOR FURTHER READING

Anonymous, *The Psychology of Getting Grades.* Columbia, Missouri: Lucas Brothers, 1935. 90 pages.

Dunlap, Knight, *Habits, Their Making and Unmaking.* New York: Liveright Publishing Corporation, 1932. 326 pages.

Forlano, George, *School Learning with Various Methods of Practice and Rewards.* New York: Bureau of Publications, Teachers College, Columbia University, 1936. Part I.

McKown, Harry C., *How to Pass a Written Examination.* New York: McGraw-Hill Book Company, 1943. 162 pages.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman & Company, 1929. Chapter XIII.

Thorndike, Edward L., and others, *The Fundamentals of Learning.* New York: Bureau of Publications, Teachers College, Columbia University, 1932. Chapters I–VII, XVI, XVIII.

---

[46] *Journal of Educational Research,* 28: 519, March, 1935.

CHAPTER XIII

Diagnosis

A. The Problem of Diagnosis in Education

The nature of educational diagnosis. Educational diagnosis seeks to determine the nature and causes of unsatisfactory adjustment to the school situation. It is concerned with the specific weaknesses of individual pupils. Diagnosis seeks not so much to describe or explain educational maladjustment as to correct or prevent it. Adequate diagnosis is the basis of all intelligent guidance and of all effective teaching.

Education borrowed the term "diagnosis" from medicine where its fundamental character has been long recognized. Medical diagnosis commonly starts with some bodily symptom, such as pain or abnormal temperature. The next step is to determine the causes that lie behind the symptoms. The trouble, may be the malfunctioning of some organ or gland, which in turn may be caused by some particular germ or toxic condition, and which, when located, may yield readily to the appropriate medical treatment or surgery. The order of events is clearly indicated by the rule: "Before you dose, diagnose!"

The situation in education is much the same although here the scope of diagnosis is usually broader. At times educational difficulties can be traced to some organic defect, such as imperfect vision or hearing, or some glandular disorder, but educational diagnosis is more often concerned with functional disorders rather than organic. Pupils who are perfectly normal organically may experience great difficulty with various aspects of the school situation. It is a matter of common knowledge that many serious learning difficulties arise, not so much from structural defects as from other factors, such as faulty habit-formation, lack of interest, or a poor home environment. Despite these complications an outstanding educator has asserted that "experts in reading, arithmetic, and spelling can now make diagnoses no less valid and reliable than are most diagnoses in medicine."[1]

Furthermore, the learning process at any time is usually condiditioned by many factors, both inside and outside the learner. It is seldom possible to isolate a single causative factor analogous to the disease germ in medicine, but the various factors may be classified roughly as follows:

---

[1] William A. Brownell, "Quantitative Research on Learning and Teaching," *School and Society,* 50: 851, December 30, 1939.

1. Internal factors:
   a. Physical: sensory equipment, glandular balance, health status, stage of maturity level, etc.
   b. Intellectual: general intelligence, specific talents and deficiencies, etc.
   c. Emotional: attitudes, interests, drives, prejudices, feelings of inadequacy, etc.
   d. Educational: background, work-habits, etc.
2. External factors:
   a. School environment: educational program, teacher, playmates, equipment, etc.
   b. Extraschool environment: home, community, church, recreational facilities, etc.

The scope of educational diagnosis has also increased to keep pace with the growing concept of education. When the conventional school conceived of its function rather narrowly in terms of certain academic knowledge and skills, the scope of diagnosis was likewise limited. Now that the modern school has enlarged the concept of education to make it synonymous with the growth of personality, it is no longer possible to limit the scope of diagnosis to locating the causes that interfere with the ordinary academic progress of the pupil. The learning difficulties presented by the school curriculum will doubtless always constitute an important part of any program of diagnosis. In fact, this phase of diagnosis naturally increases in scope and importance as the objectives of the various school subjects are extended to include the less tangible outcomes, such as attitudes, interests, appreciations, tastes, and standards of judgment. But some of the most important and difficult aspects of diagnosis have to do with social adjustments and personality disorders of many kinds.

It is likewise apparent that the scope of diagnosis is much larger than the use of tests and examinations. This, of course, does not mean that tests have an unimportant place in educational diagnosis. On the contrary, an adequate diagnosis may involve the use of intelligence tests, both general and specific, and of diagnostic achievement tests, both standardized and teacher-made, as well as the use of various pieces of laboratory apparatus for measuring sensory acuity, co-ordination, and the like. In addition to many kinds of tests, reliance must be placed upon other forms of appraisal, such as rating scales, uncontrolled observation, questionnaires, and interviews. Important as are the ordinary forms of measurement in diagnosis, they are often by themselves insufficient. Keys has well stated the role of intelligence tests in diagnosis: [2]

> Few psychologists today look to an individual's score on an intelligence test, alone and of itself, to determine the source of his difficulties or indicate the exact solution to his problems. It is entirely probable, however, that the outcome of such a test, judiciously chosen and competently administered, will contribute as much if not more to sound clinical appraisal than any other single fact obtainable. Properly supplemented with other diagnostic procedures, the information thus derived is virtually indispensable to intelligent attack upon a wide variety of problems.

The importance of educational guidance in the modern school arises from two facts: (1) many pupils make unsatisfactory progress in school—some fail altogether and others only achieve much less than the level of their capacity; and (2) few causes of maladjustment lie on the surface or are self-evident.

It should be noted that up to the present time most tests designed specifically for diagnostic purposes have been for the elementary school. As long as the secondary school and college had highly selected student bodies, their need for diagnostic tools was less acute. In recent years the enlarged enrollments at these higher levels of education have greatly increased the need for diagnosis.

The value of diagnosis in education. There is an abundance of experimental evidence to show the value of educational diagnosis combined with the appropriate remedial measures. Such evidence is available on all levels of instruction and in a variety of subjects. Science has added confirmation to the verdict of common sense: it really helps to "put the oil where the squeak is." For example, Baker [3] found that four months' special coaching of sixty nine-year-old pupils from seven Detroit schools resulted in a gain of about seven months in educational age. The coaching consisted of two thirty-minute periods per week devoted to the subject or subjects in which the pupil had shown weaknesses. Scruggs [4] compared the improvement of two equivalent classes of fifth-grade Negro children in Kansas City, one of which had the ordinary group instruction in handwriting and the other an equal amount of corrective practice based upon a detailed analysis of the weaknesses of each pupil. In seven weeks the second group increased the average quality of its handwriting about twice as much as the first. In a similar study

---

[2] Noel Keys, "Applications of Intelligence Testing," *Review of Educational Research,* 8: 256, June, 1938.

[3] Harry J. Baker, *Educational Disability and Case Studies in Remedial Teaching,* page 65. Bloomington, Illinois: Public School Publishing Company, 1929.

[4] Sherman D. Scruggs, "Remedial Teaching for Improvement in Handwriting," *Journal of Educational Research,* 23: 288–295, April, 1931.

Guiler[5] found that fourteen seventh-grade pupils made in three months a normal gain of three years in quality of handwriting. Blair[6] has summarized studies in the tool subjects on the secondary level which show similar results.

It has been shown that the value of such remedial measures is by no means confined to skill subjects, such as handwriting and spelling. For example, a study by Leonard[7] showed that junior high-school pupils improved more rapidly in the ability to write compositions free from common errors in capitalization and punctuation during a program involving error analysis and appropriate remedial exercises than did pupils of like ability exposed to the conventional method of teaching. While both groups showed definite improvement, the mean decrease in the twenty-eight most frequent errors, after eleven forty-five-minute practice periods, was approximately twice as great for the experimental as for the control groups. Experiments by Guiler[8] on the elementary-school, the senior-high-school, and the college levels showed comparable results from similar methods.

Stone[9] found that pupils in the fifth and sixth grades in twenty-three schools, who devoted not more than forty minutes a day for five weeks to diagnostic and practice tests, gained two to six times as much in ability to solve reasoning problems as did pupils of equal ability who had only the regular arithmetic work in school. Furthermore, the results of the study indicated that the superior gain in reasoning ability resulting from this diagnostic and remedial program was about twice as great for pupils in the highest sixth in intelligence as for those in the lowest sixth, that the gain transferred to problems of a different content, and that it persisted for at least a year, at the end of which the retests were given.

The psychology of such procedures seems reasonably clear. It is a sound principle of teaching which holds that learning always begins where the learner's present knowledge leaves off. Failure to observe this principle results in foolish attempts to do two impossible things. One of these is attempting to teach a pupil what he already knows. The other is attempting to teach him on a level too far beyond his present knowledge. Both are equally futile. The only adequate safeguard to be obtained is in frequent check-ups on the pupil's progress.

---

[5] Walter Scribner Guiler, "Improving Handwriting Ability," *Elementary School Journal,* 30: 56–62, September, 1929.

[6] Glenn Myers Blair, *Diagnostic and Remedial Teaching in Secondary Schools,* 422 pages. New York: The Macmillan Company, 1946.

[7] J. Paul Leonard, "The Use of Practice Exercises in Teaching Capitalization and Punctuation," *Journal of Educational Research,* 21: 186–190, March, 1930.

[8] See *Elementary School Journal,* 34: 427–437, February, 1934; *School Review,* 41: 450–458, June, 1933; and *Educational Research,* 26: 110–115, October, 1932.

[9] C. W. Stone, "An Experimental Study in Improving Ability to Reason in Arithmetic," *Twenty-Ninth Yearbook of the National Society for the Study of Education,* pages 589–599. Bloomington, Illinois: Public School Publishing Company, 1930.

B. The Techniques of Diagnosis

The levels of diagnosis. The process of educational diagnosis may be profitably thought of as falling into five steps, or levels.

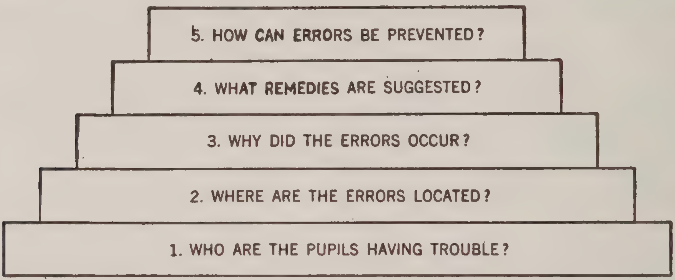

5. HOW CAN ERRORS BE PREVENTED?
4. WHAT REMEDIES ARE SUGGESTED?
3. WHY DID THE ERRORS OCCUR?
2. WHERE ARE THE ERRORS LOCATED?
1. WHO ARE THE PUPILS HAVING TROUBLE?

Figure 41. The Five Levels of Educational Diagnosis.

Figure 41 is a graphical representation of the process. It will be noted from the questions asked at each level that the first four steps —the *W*'s—have to do with corrective diagnosis, while the highest level has to do with what may be termed preventive diagnosis. In other words, the immediate purpose is *correction,* but the ultimate purpose is *prevention.*

Locating the individuals needing diagnosis. How can we best locate the pupils not making satisfactory adjustment to the school situation? This is logically the problem with which the program of educational diagnosis begins. The order of events is not unlike that described in the famous recipe for making rabbit stew which begins: "First you catch your rabbit." Strictly speaking, however, while it is a necessary preliminary step, it is hardly a part of the actual process of diagnosis.

Various ways of locating the individuals who require diagnostic study have been used. Survey and group intelligence tests are often employed to screen those whose achievement is unsatisfactory. Using this method Wilson[10] found that about 70 per cent of the pupils in the seventh and eighth grades of fifteen representative cities and towns in the metropolitan area of Boston needed corrective instruction in the fundamental arithmetic processes.

Several writers suggest that any pupil whose level of achievement is well below his level of intelligence is worthy of special study. Others contend that a practical difficulty with the procedure is that tests of achievement and so-called tests of intelligence really largely measure the same thing, and suggest instead that diagnostic study be given to those pupils whose achievement in some school subject, or subjects, is well below their general achievement level. Still other writers rely heavily upon the judgment of the teachers. Baker,[11] for example, selected his sixty pupils for special remedial coaching by taking those who had received final marks of failure or conditional passing in four fundamental subjects. He admits that this criterion was used at the outset primarily because of its availability, but states that it "arose steadily in our esteem."

---

[10] See *The Role of Research in Educational Progress,* pages 234–241. Washington: American Educational Research Association, 1937.

All these suggestions have merit. The judgment of the present teacher should always be taken into account, especially since in the ordinary school whatever diagnostic and remedial work is attempted will be undertaken by the regular classroom teacher. But the present teacher's judgment needs to be supplemented by considering the judgment of past teachers as reflected in the school record. Since the judgment of teachers is not infallible, however, general achievement tests and intelligence tests will be found particularly valuable. Any pupils in the intermediate grades whose achievement falls a year or more below their age or grade level should usually merit special study. Discrepancies between achievement and intelligence are of particular significance when intelligence has been measured by individual tests or performance tests rather than by ordinary group tests. Such discrepancies also assume added significance when the pupil has apparently had ample opportunity for learning.

While special study and treatment are often justified for the lowest 5 or 10 per cent in the typical class, it must not be thought that diagnosis should be restricted to low-ranking pupils and to obvious misfits. On the contrary, some of the most profitable cases are those whose achievement is average or even above, but is nevertheless well below what appears possible. As a matter of fact, Hildreth [12] points out that many clinics prefer not to attempt remedial work with very dull pupils, say those with IQ's of approximately 80 and below, but prefer instead to alter the achievement goals for such children. It will be found at times that pupils whose personality defects interfere with satisfactory social adjustment represent superior academic achievement. In fact, psychiatrists point out that the teacher should often be most concerned about the mental health of those who give her least concern academically. The writer recalls the case of a sixth-grade girl whose scholastic achievement was well above the norms on the tests but whose attempts at social adjustment to the group had been distinctly unsuccessful. The girl told her mother that she would give anything in the world if she had just one friend. In the conventional school this girl would have been regarded as making an entirely satisfactory record, but in the modern school she is seen to be so seriously maladjusted as to require special treatment.

---

[11] Harry J. Baker, *op. cit.*, pages 9–16.

[12] Gertrude Hildreth, *Learning the Three R's* (Second Edition), page 545. Philadelphia: Educational Publishers, Inc., 1947

Locating the nature of the difficulty. After locating the pupils who are experiencing trouble, the next step is to make a careful examination of the difficulty of each pupil. A bill of particulars is needed. It is just here that diagnostic tests, if available, are of great value. The aim of such tests is to reveal the specific location of the pupil's difficulties. As a rule, each test has a limited scope, but attempts to explore thoroughly this restricted area. For example, one test might undertake to find the particular number combinations which are causing trouble in the addition of whole numbers, while another test attempts to find out whether inadequate reading ability, faulty technique of analysis and procedure, lack of skill in the fundamental processes, or some other factor is responsible for poor performance in reasoning problems.

Most of the diagnostic tests published to date are limited to the tool subjects mainly on the elementary level. Traxler [13] has prepared a comprehensive bibliography of available tests together with a practical discussion of their effective use. Blair [14] has compiled similar information with special reference to the high school. Recently Traxler [15] offered this warning: "Our experience at the Educational Records Bureau indicates that, at present, there is scarcely one test which gives us as much reliable information as is needed for effective diagnosis in any one field."

But any test, whether standardized or not, can be used to reveal the location of errors. The principal advantages of the standardized test are that in content it is likely to represent a more careful selection than the informal test, and that the existence of comparable forms makes it possible to verify the accuracy of diagnosis based on one form and to check upon the success of any remedial measures undertaken. However, these special values in standardized tests by no means rule out the values of informal tests when used for diagnostic purposes.[16] In reading, for example, some writers regard informal tests as even more important than standardized tests. The diagnostic value to be realized depends more upon the teacher than upon the test used. Durrell estimates that at least 75 per cent of the cases requiring special attention in reading can be handled adequately by well-trained classroom teachers using non-standardized tests supplemented by observation of the pupils' achievement and work habits. He says:

---

[13] Arthur E. Traxler, *The Use of Test Results in Diagnosis and Instruction in the Tool Subjects,* 80 pages. New York: Educational Records Bureau, 1942.

[14] Glenn Myers Blair, *op. cit.*

[15] Arthur E. Traxler, "Individual Evaluation," in *New Directions for Measurement and Guidance,* page 28. Washington: American Council on Education, 1944.

> Such informal tests and observation charts usually indicate the correct level on which to start remedial instruction, the specific reading abilities in which the child is weak, and the faulty habits and confusions which must be overcome in the remedial program.[17]

Figure 42 illustrates a useful procedure for analyzing the errors revealed by a standard test in arithmetic. The procedure is equally applicable to informal tests. This particular test, Test 3, Arithmetic Fundamentals, of the Metropolitan Achievement Tests, was administered to a fifth grade in October. The pupils are arranged in descending order according to the score on this test. Each error is indicated by X and each omission by 0 as far as the pupil attempted problems; the problems beyond the last one attempted are indicated by - - - - . The summary at the bottom shows how many times each problem in the test was missed and omitted. This simple analysis reveals clearly what type of problems caused trouble and to whom the trouble was caused. The procedure is really group diagnosis, but it may be regarded as the first step in individual diagnosis. It should be apparent that classroom teachers who are content merely to obtain the total score made by each pupil on a test are really overlooking the greatest value of the test for instructional purposes.

Similar error analyses can be made for most subjects, but are especially valuable in mathematics, spelling, reading, handwriting, and language. It is usually better to make more than one such analysis, however, than to rely upon a single sampling, which is almost sure to include some errors that are merely chance occurrences rather than habitual. Brueckner and Elwell,[18] for example, found from the study of a test in the multiplication of fractions, containing in random order four examples of each type, that failure to work a single example correctly is hardly a safe index and that

---

[16] Donald D. Durrell, *Improvement of Basic Reading Abilities,* page 18. Yonkers: World Book Company, 1940.

[17] *Ibid.,* page 296. Quoted by special permission.

[18] Leo J. Brueckner and Mary Elwell, "Reliability of Diagnosis of Error in Multiplication of Fractions," *Journal of Educational Research,* 26: 175–185, November, 1932.

NAME OF PUPIL	WHOLE NUMBERS																					FRACTIONS																	
	Addition					*Subtraction*					*Multiplication*					*Division*						*Addition*					*Subtraction*									*Division*			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
Peggy					X									X							X		X	X	O	O		X	X	O	O		O	O	O		O	O	O
Mildred				X	X										X				X	X	X			X	O	O				O	O		X	O	O	O	O	O	O
James					X										X			X	X		X		X	X	O	O	X	X	O	O	O		X	O	O	O	O	O	O
Betty W				X	X																	X	X		X	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Billy															X		X						X	X	X	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Dorothy					X										X								X	X	X	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Esther									X					X			X	X			X		X	X	O	X		X	X	X	X		X	O	X	X	X	X	X
Ruth D				X	X									X	X		X	X	X		O		X	X	X	O			X	X	X		X	O	O		O	X	O
Mary					X								X	X					X		X	X		X	X	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Betty B														X	X		X	X	X	X	X		X	X	X	X		X	X	O	X	—	—	—	—	—	—	—	—
Bobby				X	X					X				X	X						X		X	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Ann				X	X									X	X						O	X	X	X	X	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Jeanne					X									X	X		X	X	X				X	X	X	X	—	—	—	—	—	—	—	—	—	—	—	—	—
Nancy								X	X	X					X	X		X			X		X	X	O	O	X	O	X	—	—	—	—	—	—	—	—	—	—
Marybell				X	X									X	X				X	X	O		X	X	X	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Ruth L				X	X									X	X				X	X	X		X	X	X	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Howard					X									X	X		X	X	X	X		X	X	X	X	—	—	—	—	—	—	—	—	—	—	—	—	—	—
John				X	X	X	X	X	X	X				X				X	X	X			X	X	X	X	O	O	O	O	O	O	O	O	O	O	O	O	O
Ben				X	X			X					X	X	X		X	X	X	O	O		X	X	X	O		X	X	O	X		X	O	O	X	O	O	O
Lewis				X	X										X		X	X	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Frazer								X	X	X			X	O	X		X	X			X		X	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Ewing				X	X					X				O	O		X		X	O	O		X	X	X	X	X	X	X	X	—	—	—	—	—	—	—	—	—
Julia				X	X			X		X			X	X	X				X	X	X		X	X	X	X	X	X	X	X	X	X	X	—	—	—	—	—	—
Nelle				X	X	X			X	X				X	X		X	O	X		X		X	X	O	O	X	—	—	—	—	—	—	—	—	—	—	—	—
May					X		X		X				X	X	X	X	X	X		X	X		X	X	X	X	—	—	—	—	—	—	—	—	—	—	—	—	—
Dick					X				X	O		O	O	O	O			O	O	O	O	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Sally					X			X	X	X				X	O		X	X	X	X	O	X	X	X	X	O	X	—	—	—	—	—	—	—	—	—	—	—	—
Sam					X			O	O	O		X	X	O	O			X	X	O	O	O	O	X	X	O	O	O	O	O	O	O	O	O	O	O	O	O	O
SUMMARY																																							
No. Errors				13	23	2	2	6	8	8		1	6	17	19	2	13	14	16	9	13	5	24	23	18	7	6	7	8	4	5	1	6		1	2	1	2	1
No. Omissions								1	1	2		1	1	4	4			2	2	5	9	3	3	4	10	21	16	19	19	24	23	21	22	28	27	24	27	26	27

Figure 42. Analysis Sheet of Test 3, Metropolitan Achievement Tests, Form A, Arithmetic Fundamentals, for a Fifth-Grade Class in October (last 21 problems are omitted).

at least three problems of each type are required for a valid individual diagnosis. A later study[19] in subtraction showed that all the problems of a type should be grouped together on the test.

It is not sufficient, however, to stop with tabulating the frequencies of questions missed on tests or mistakes made in written work. A further analysis must be made of the types of errors represented. It will be noted that problem 24 in Figure 42 was missed by 23 pupils out of 28. As a basis for remedial instruction the teacher needs to know what types of incorrect solutions were made by her pupils. An examination of the test papers provides the answer. Problem 24 follows:

24. Add

$$\begin{array}{r} 3\frac{1}{3} \\ \underline{4\frac{1}{6}} \\ \boxed{} \end{array}$$

It is found that 15 of the 23 incorrect solutions were $7\frac{3}{6}$, merely a failure to reduce the fraction to its lowest terms. Five of the 6 errors made by the best 7 pupils were of this type. Five pupils got as an answer $7\frac{2}{6}$, which represents two types of errors. Still more serious is the status of the pupil who got $\frac{7}{6}$ for an answer. An interesting type of incorrect solution is represented by a pupil whose answer was $7\frac{2}{9}$. It is apparent that he merely added the numerators and the denominators without taking the trouble to reduce the fractions to a common denominator. The other wrong answer was $7\frac{2}{3}$.

A second illustration of the value of error analysis is taken from spelling. A few years ago the writer gave a spelling test to a class of high-school seniors. The results were disappointing. One of the words missed most often was "undoubtedly." Contrary to expectation, a tabulation of the errors revealed the fact that the first two syllables were spelled correctly by all pupils. The misspellings were of four forms: "undoubtelly," "undoubtely," "undoubtaly," and "undoubtally." It can be seen that the fundamental error is mispronunciation. The pupils were attempting to *spell* this common word as they were accustomed to *pronounce* it. Hildreth[20] reports that confusion over vowels in the middle and end syllables is a prolific source of error, and that syllables containing *e, a,* and *o,* are especially liable to vague, indistinct pronunciation. Another investigator[21] found that emphasis upon correct pronunciation in reading resulted in a decided improvement in the spelling of pupils in the fifth and sixth grades.

---

[19] Leo J. Brueckner and Mabel J. Hawkinson, "The Optimum Order of Arrangement of Items in a Diagnostic Test," *Elementary School Journal,* 34: 351–357, January, 1934

[20] Gertrude Hildreth, *op. cit.,* page 492.

One of the greatest values of such error analyses is that they reveal that a relatively few types of errors made over and over again are responsible for the poor performance of most pupils. In an early study of errors in spoken language Charters[22] found that 71 per cent of the errors made by Pittsburgh children fell into only five classes. A more recent study in Madison, Wisconsin,[23] revealed that more than half of the total number of language errors made from the kindergarten through the sixth grade represented but four types. In an extensive study Newland[24] found that errors in writing only four letters, *a, e, r,* and *t,* accounted for almost half of the illegibilities made, whether by elementary-school, high-school, or adult groups, and that only four types of difficulties in letter-formation caused more than half of the illegibilities. It cannot fail to be encouraging to teachers and pupils alike to find that remedial efforts directed at a relatively few troublesome points will result in great improvement.

It has also been found that serious emotional disturbances and antisocial conduct often result from some educational deficiency. Stullken,[25] for example, found that about 25 per cent of the pupils of the Montefiore School in Chicago, a special school for poorly adjusted boys, had distinct reading disabilities and that when these were corrected the boys often ceased to be problem cases.

Locating the causes of errors. Even more important, and usually far more difficult, than knowing *where* the errors occur is knowing *why* they occur. One limitation of test scores in diagnosis is that they reveal the *products* of learning rather than the learning *process* itself. Tyler[26] makes a useful distinction between measurement or appraisal, and interpretation or inference. In other words, causation is not established directly by the act of measurement, but must be inferred from the measurement and other pertinent data. Scates[27] puts the situation clearly:

---

[21] Marjorie E. Kay, "The Effect of Errors in Pronunciation upon Spelling," *Elementary English Review,* 7: 64–66, March, 1930.

[22] Unpublished report made in 1919.

[23] *Language Curriculum, Committee Reports.* Madison, Wisconsin: Madison Public Schools, 1932.

[24] T. Ernest Newland, "An Analytical Study of the Development of Illegibilities in Handwriting from the Lower Grades to Adulthood," *Journal of Educational Research,* 26: 249–258, December, 1932.

[25] Edward H. Stullken, "The Philosophy of the Special School," *Phi Delta Kappan,* 22: 345–350, March, 1940.

[26] Ralph W. Tyler, "Elements of Diagnosis," *Thirty-Fourth Yearbook of the National Society for the Study of Education,* page 113. Quoted by permission of the Society. Bloomington, Illinois: Public School Publishing Company, 1935.

> A multitude of test scores are in themselves meaningless. They show facts, but they do not show reasons. They neither diagnose nor evaluate. They may be useful aids, but they leave the principal problem to the teacher's insight, namely, that of determining what is indicated.

At times, as in some of the examples cited, a reasonably safe inference can be made from the nature of the errors themselves. But rarely can a sufficiently complete explanation be made without considering the child's past history, outside the school as well as inside. It is never safe to infer that a child's poor performance in school is due to mental deficiency or laziness or other personality defects, unless a careful study of his educational opportunities has been made. Fortunate indeed is the school whose records are sufficiently complete to provide the essential data.

Certain outstanding physicians and surgeons have advocated an enlarged concept of diagnosis in modern medicine. Several years ago Sir William Osler argued that it was more important to know what kind of man had a certain disease than to know what disease the man had. The president of Stanford University has made the following statement:[28]

> It is just as important in these days for a young doctor to understand his patient's personal life, home responsibilities, and community relationships, as it is to be able to tell just what organisms are living in his lungs or invading his liver. . . . The doctor who has not studied psychology and who cannot acquire a knowledge of it, if he is to be successful, will have to confine himself to work in the laboratory or be a pure technician.

Hildreth suggests that the following five "areas of investigation"[29] are important in diagnosis:

> *Mental equipment of the learner.* Aptitude for academic schoolwork, learning capacity, readiness for learning, habitual modes of response, judgment, reasoning ability, insight, memory, association, perception, attention span, ability to see relationships, creative ability, intellectual interest, suggestibility, comprehension, auto-criticism, habits.
>
> Language equipment: Command of mother tongue, knowledge of foreign languages, language first learned, speech defect, immaturity in speech, in articulation, or diction; vocabulary, rapidity or slowness of speech, history of speech development, age of using words and sentences, descriptive powers, written composition.

---

[27] Douglas E. Scates, "Differences Between Measurement Criteria of Pure Science and of Classroom Teachers," *Journal of Educational Research,* 37: 1–13, September, 1943.

[28] Ray Lyman Wilbur, "The March of Medicine," *Science,* 87: 201–202, March 4, 1938.

[29] Gertrude Hildreth, *Learning the Three R's,* pages 547–549. Minneapolis: Educational Publishers, Inc., 1936.

Personality, temperament, and dynamic equipment. Self-control, affability, desirable and undesirable inhibitions, attitudes, friendliness, susceptibility, docility, irascibility, drive, perseverance, stability, lability of mood, compliance, responsiveness, restlessness, shyness, tendency toward embarrassment, day dreaming, fears, withdrawal from reality, sex interest, morbid curiosity, irrational attitude, manners, attitude toward failure and toward the school disability, compensations, child's interests, attitude toward school, preferred school subjects, child's play interests, obsessions, fears, worries, ability to get along with other children, social qualities, attitude toward brothers and sisters and other members of the family, delinquent and anti-social activities, degree of normal adjustment, changes, growth and development in all these factors since birth.

Physical status, sensory and motor equipment, physical conditions. Sensory acuity, constitutional defects, physical maturation, physical handicaps and defects, disease history, glandular balance, condition of teeth, etiology of illness, posture, accidents or unusual physical shocks, nutrition, diet, hygiene, psycho-motor status, muscular strength or weakness, handedness, steadiness, coordination, efforts to change handedness, facility in sports and games.

Environment and home history. Economic factors, literacy of parents, number of sibs, marital status of parents, foreign background, other adults in the home and their contact with this child, evidences of culture, e.g., books, musical instruments, labor-saving devices in the home, harmony in home adjustments, attitude of home toward school, cooperation of home with school, neighborhood environment, association with other children, child's opportunity for free time, child's activities in free time.

Child's daily schedule: Rising, eating, sleeping, play, schoolwork at home; regularity or irregularity in home program.

School situation, history and present status. Methods of instruction, especially in the work with which the child has difficulty; size of class groups, capability of class groups, school marks, textbooks and other materials used, progress of other children, progress in learning from grade to grade, retardation, failure or double promotion, attitude of child toward teacher, teacher's usual success with pupils of her grade level, teacher's experience, rapidity with which average child progresses, requirements of the course of study, classification system, provision for individual assistance, date of first recognition of the child's disability, former diagnostic and remedial work carried on with the child both in school and in clinics, survey of all school records that would throw light on the situation, kind and extent of supervision, objective test records, analysis of previous training and methods of attack in learning, e.g., to write; evidence of readiness for instruction before work in skills, teacher's story of the case, attitude toward the child, discipline in the classroom, absence from school, tendency toward tardiness, truancy. Information the teacher has about modern methods in education and child study, progressiveness of the school program, extent to which teacher makes individual studies and keeps cumulative records of pupil, age of the child on entering school, terms retarded, failure in specific subjects, absence. Extent to which each teacher is acquainted with the child's past school history, extent to which the teacher knows the facts at the beginning of the school term, teacher's explanation of the cause of difficulty, teacher's recommendations as to what should be done, efforts the teacher has been making to eradicate the difficulty, extent to which the teacher capitalizes the child's interests.

It is, of course, manifestly impossible, as well as usually unnecessary, to consider all these facts in any particular case. Satisfactory explanation of the less serious cases can often be found in a relatively few factors, although rarely ever in just one. The more serious cases will usually be found more complex to analyze as well as more difficult to remedy.

It will frequently be necessary to supplement the data of the existing school records. A visit to the pupil's home is often helpful. A careful observation of the pupil at work is another fruitful source of information. Objective records of observations made under controlled conditions are particularly important. Considerable light is often thrown upon the attitudes and work habits of unsuccessful pupils by observing them at work and then by comparing successful pupils under similar conditions.

A skillful interview by a tactful teacher will sometimes give a clue to the difficulty when other methods fail. In the upper grades and the high school, check lists, questionnaires, and other forms of written responses are valuable aids to the personal interview. Having the pupil "think out loud" through the solution of a problem in mathematics or science, or give an explanation of the procedure used, is often most illuminating.

Two illustrations make clear the value of the interview as a supplement to the written test in locating the sources of difficulty in arithmetic. Buswell tells of a boy of better than average intelligence whose work in column addition was both slow and inaccurate. To the interviewer he explained that he did not like to add and so wanted to get the worst of it over as soon as possible. For this reason he always added the numbers according to size, beginning first with the largest numbers and leaving the smallest ones till last. But as this technique meant skipping up and down the column, it involved great risk of omitting some of the numbers altogether and of adding others more than once. The story is told of a sixth-grade school-girl who had an elaborate but usually ineffective "system" for solving reasoning problems. Her explanation was somewhat like this: "Whenever there's lots of numbers, I add, but when there's only two numbers with lots of parts [digits], I subtract. But if there is just two numbers and one is littler than the other, I divide when they come out even, and multiply when they don't." It is most unlikely that any analysis of test papers, or observation of the pupils at work, would have resulted in a correct inference as to the real trouble in either of the cases above.

Teachers often find that an interview with the pupil sheds needed light upon difficulties in reading and English. Pressey and Campbell[30] report that one ninth-grade pupil explained capitalizing the word "Pirates" on the ground that pirates are real persons just as much as "John Silver" or "Captain Kidd." Another teacher discovered that a boy had written "a quarter to three" in answer to a question on a reading test when the correct answer was "twenty-five minutes till three" because everybody knows that twenty-five cents make a quarter!

Brownell[31] has shown the possibilities of classifying the mental processes used by the pupils as revealed by interviews according to levels of maturity represented. He concludes that a reasonably flexible interview technique in analyzing learning is "exceedingly valuable if it is sagaciously employed." One survey[32] of the experimental literature relating to the reliability of the interview arrives at the conclusion that "with well-trained interviewers working under carefully defined conditions, quantitative interview ratings representing a complex over-all evaluation can be made as reliable as most personality tests, and more reliable than some of them." Nevertheless good interviewing requires skill as well as time and patience.

Remedial procedures. The ultimate purpose of diagnosis is to afford a basis for effective remedial procedures. When the cause or causes of the pupil's unsatisfactory adjustments have been dedetermined, an intelligent program of correction can be planned, and not until then. Whenever the same causes appear to operate in several pupils, group measures will be satisfactory. Usually, however, remedial programs must be planned for each pupil individually.

A study by Davis[33] shows the close relationship between educational diagnosis and remedial instruction. Two extra periods a week were devoted to 275 pupils of poor spelling ability in grades 2B to 6A, inclusive. The results showed "marked improvement." Pupils remained in the remedial classes until they made perfect scores on the spelling tests of two successive Fridays. The average time required was 7.5 hours, and bore little relationship either to intelligence or grade location. Twenty-four different types of difficulties were located, and listed with each difficulty were the most successful remedies found by the teachers. The ten most common difficulties, with their remedies, are shown in Table 33.

---

[30] Sidney L. Pressey and Pera Campbell, "The Causes of Children's Errors in Capitalization: A Psychological Analysis," *English Journal,* 22: 197–201, March, 1933.

[31] William A. Brownell, "Rate, Accuracy and Process in Learning," *Journal of Educational Psychology,* 35: 321–327, September, 1944.

[32] Sidney H. Newman, Joseph M. Bobbitt, and Dale C. Cameron, "The Reliability of the Interview Method in an Officer Candidate Evaluation Program," *American Psychologist,* 1: 103–109, April, 1946.

[33] Georgia Davis, "Remedial Work in Spelling," *Elementary School Journal,* 27: 615–626, April, 1927.

TABLE 33

DISTRIBUTION OF SPELLING DIFFICULTIES AND SUCCESSFUL REMEDIES (AFTER DAVIS)

Difficulties and Remedies	*Frequency*
1. Has not mastered the steps in learning to spell a word	88
a. Teach steps until every child knows them and uses them.	
b. Study each word with the children.	
2. Writes poorly	88
a. Discover particular letters or combinations of letters that are difficult and practice on these letter combinations.	
b. Practice words containing writing difficulties.	
3. Cannot pronounce the words being studied	78
a. Go over the words before the children study them so that every child will know what he is studying.	
b. Help the child to unlock words for himself.	
4. Has bad attitude toward spelling	71
a. Supervise study closely so that the child will get into the habit of studying words correctly without wasting time.	
b. Try to show need for study.	
c. Give study work under time pressure.	
d. Try to appeal to pride.	
e. Try to work up competition with self (that is, of the pupil with himself).	
f. Give reward.	
5. Does not associate the sound of the letters or the syllables with the spelling of the word	49
a. Teach letter sounds.	
b. Listen to careful pronunciation.	
c. Teach the child to syllabify words.	
d. Say words slowly again and again to hear sounds.	
6. Needs more time than can be devoted to spelling in the regular class	21
a. Give more time after school or during the day when other work is finished.	
7. Is discouraged because he misspelled so many words in the Monday test	20
a. Take a few words at a time.	
b. Study at odd times during the day.	
c. Have the pupil stay longer in the afternoon than the others.	
8. Has speech defect	16
a. Listen to pronunciation.	
b. Look at word carefully.	
c. Teach difficult combinations.	
9. Does not mark paper correctly	16
a. Teach child how to check.	
b. Insist on rechecking.	
c. Always check paper.	
10. Interchanges letters	10
a. Study words carefully.	
b. Underline difficult part.	
c. Try to spell by syllables.	

Traxler[34] has prepared some very convenient charts which outline appropriate diagnostic and remedial procedures for common types of disabilities in reading, arithmetic, language usage, spelling, and handwriting. Figure 42 shows the chart for handwriting. Note that a detailed analysis of samples of the pupil's writing is suggested, as well as diagnostic charts and tests.

CHART V: HANDWRITING

SUGGESTED DIAGNOSTIC AND REMEDIAL PROCEDURES

TYPE OF DEFECT	DIAGNOSTIC PROCEDURE	SUGGESTED TYPES OF REMEDIAL TREATMENT
1. Slant a. Too much slant b. Writing too straight c. Lack of uniformity	1. Use diagnostic chart; study different samples of writing. Draw lines through letters parallel to slant on different parts of page. Compare these lines as to direction. Observe pupil as he writes and note details—position, paper, etc.	1. Some instances of poor slant can be corrected by changing position of writing arm or manner of grasping pen. Change in position of paper will help others. Note that paper should be at an angle. Other pupils must learn to turn their hand as they approach end of line. Explain to pupils effect of slant on quality.
2. Alignment a. Lack of uniformity b. All letters about the same height	2. Use diagnostic chart; draw horizontal lines through writing even with top and bottom of some of the letters.	2. Explain defect to pupil. Lack of uniformity of alignment results partly from motor inco-ordination and will probably be corrected as co-ordination of writing movements improve through practice.
3. Quality of line a. Writing too heavy b. Writing too light c. Line wavy and uncertain	3. Use diagnostic chart; note type and size of pen and manner of holding it; note speed of writing.	3. Make sure that pupil has proper writing materials; see that he does not use his writing arm to support his body. If line is thin and wavering, give drills to speed up movement and improve co-ordination.
4. Formation of letters a. Poor general form b. Lack of smoothness c. Parts omitted d. Parts added e. Letters not closed	4. Use diagnostic chart; if desired, letter form may be analyzed in detail with Pressey chart. Study general form and habits of forming each letter. Often faults in letter form are related to only a few letters.	4. Make some use of movement drills to improve smoothness. Practice especially on movements common to several letters. Study details of letter form with pupils and show them where they need to improve. Have pupils practice individually on the letters which diagnosis has shown to be poorly formed.

---

[34] Arthur E. Traxler, *op. cit.*, pages 34–35.

CHART V (Continued)

Type of Defect	Diagnostic Procedure	Suggested Types of Remedial Treatment
5. Spacing of words a. Too wide b. Too narrow c. Not uniform 6. Spacing of letters a. Too wide b. Too narrow c. Not uniform	5. and 6. Use diagnostic chart; study various samples; note whether wide spacing or crowding occurs on different part of page. Observe pupils while writing for evidence of too much lateral movement.	5. and 6. Explain fault to pupil. Have him pay especial attention to spacing while writing samples to be inspected by teacher. Movement exercises are of some value in improving spacing.
7. Size of writing a. Too large b. Too small c. Lack of uniformity	7. Study different samples and compare with those of other pupils in the same grade. Considerable variability is allowable among individuals and especially between grades. Young pupils tend to write large. Note freedom of movement. Try to discover cases of lack of uniformity.	7. Writing that is too small may result from a cramped finger movement. Give movement exercises to relax pupil and bring about some arm movement. If writing is too large pupil can sometimes correct it through conscious effort if his attention is called to it. In young pupils, improvement may have to await the process of maturation.
8. Writing not neat a. Blotches b. Words crossed out and rewritten	8. Examine samples of writing, especially those prepared in daily work. With respect to blotches see if writing materials are defective.	8. See that pupil has proper writing materials and that they are kept in working order. Explain effect of lack of neatness on all school work. Make daily work in other subjects the gauge of neatness.
9. Speed a. Writing too slow b. Writing too fast	9. Speed of writing affects the quality, but aside from this fact it is important in that some pupils write so slowly and laboriously that they have difficulty in preparing assignments on time. Give a test of speed of writing and compare number of letters per minute with grade norms.	9. If writing is too fast show pupil its effect on letter form and have him write samples under timed conditions. Some pupils go to the other extreme and write so slowly that they practically draw the letters. Give movement exercises while counting rapidly to break down habits of slow movement. Insist that pupils speed up writing regardless of their letter forms. Their writing will probably deteriorate for a time, but when old habits are broken down teacher and pupil can build new ones.

Figure 43. Traxler Chart of Suggested Diagnostic and Remedial Procedures in Handwriting.

An effective method with bright pupils may fail with dull. In fact, no method is likely to improve materially the academic achievement of the mentally deficient child. Even with normal or superior children the substitution of correct habits for incorrect will require time. No sudden transformation is to be expected. But if only negligible progress results from extended practice, the remedial program should be revised.

Preventive diagnosis. In the long run, the greatest value of a diagnostic and remedial program is the discovery of preventable factors within the control of the school which lead to maladjustment. Frequently, modifications in school organization, curriculum, instructional materials, and teaching methods are suggested by an analysis of what is happening to the pupils under the existing program. Manifestly, factors which have produced learning difficulties in the past are likely to do so in the future. It is always better, and generally easier, to prevent errors than to correct them. It will often be found that a program of studies which provides wider differentiation in method and content to suit pupils of varying abilities and interests is the way out of many difficulties. The systematic use of readiness tests of various types to determine when the pupil is sufficiently mature, physically, mentally, and socially, to begin the regular work of the first grade, and a judicious use of aptitude tests to establish the pupil's fitness for the more formal and abstract subjects, such as arithmetic, algebra, and foreign language, will prevent much needless failure. In the foreword to a recent book on remedial teaching Terman [35] says: "Perhaps the most important conclusion to be drawn from the extensive researches here reported is that disability of any degree in any of the basic school subjects is wholly preventable." *Prevention is the highest level of diagnosis, its ultimate goal.*

C. Diagnosis in Reading

Diagnostic and remedial work in reading affords excellent illustrations of the above techniques. No subject in the curriculum has received more attention in recent years or shown greater development than reading. Diagnostic and remedial work in this subject has demonstrated its effectiveness on all levels of instruction from the first grade to college. The space available here is sufficient only to point out the more important relationships of tests and measurements of various types to the reading program of the modern school. The discussion is organized under two headings, as follows:

---

[35] Lewis M. Terman, "Foreword" to Grace M. Fernald's *Remedial Teaching in Basic School Subjects,* page ix, copyright 1943. Reprinted by permission of the publishers, McGraw-Hill Book Company, Inc.

1. Some representative case studies.
2. The general technique of diagnostic and remedial work in reading.

1. Some Representative Case Studies

CASE NO. 7,[36] BOY, GRADE 3A, CA 10–8

A. *Nature of the Difficulty (Symptomatic Behavior).*

1. Social:
    - a. Very shy. Spends much time in daydreaming.
    - b. Complies readily with any request made of him. Loath to make choices, saying, "It doesn't make any difference to me," or "I don't care."
    - c. Has many playmates who seem to like him.
    - d. Greatly embarrassed because children laugh at him because he cannot read.
2. Educational:
    - a. So anxious to learn to read that his very eagerness inhibits him. Gets a book and tries to read whenever he has leisure time.
    - b. Tries to "sound out" words but is not successful. Does not know sounds of consonants.
    - c. Repeats words and phrases excessively.
    - d. Confuses letters which look alike—*b* and *h, e* and *c, t* and *k.*
    - e. Often loses his place but goes right on as though he knew where he was, making up the story.
    - f. Sometimes remembers a story verbatim after one faulty reading of it.
    - g. Holds book close to his face, then moves it to a distance, or vice versa.
    - h. Test scores in silent reading about 1A level. Makes no score in oral reading. Record in other subjects equally poor.
    - i. "Sometimes he knows a word, and the next day he doesn't."

B. *Causative Factors.*

1. Physical:
    - a. Visual word images were confused and blurred by defective vision. Ten months ago both eyes tested 20/30; now they are 20/70.
    - b. Auditory word images were similarly confused and blurred by defective hearing. Both ears 20/30; variable, better some days than others. Has severe earache. Running ears since tonsils removed at age 5.
    - c. Improper breathing, defective teeth, and earache make him listless and uncomfortable, and therefore incapable of well-directed attention.
2. Intellectual:
    - a. Has MA of 9 years, and IQ of 83, on the tests. Examiner felt he was more intelligent than test results would indicate; probably normal. However, teachers felt he was not intelligent, since he had not learned to read.
3. Educational:
    - a. Has attended one school regularly.
    - b. Spent four terms in 2A.

---

[36] Adapted from Harry J. Baker and Bernice Leland, *In Behalf of Non-Readers,* pages 19–24. Bloomington, Illinois: Public School Publishing Company, 1934.

  - c. Much "extra" help in school, but to no avail.
  - d. Principal refuses to accept the idea that defects in vision and hearing bear a vital relation to his inability to learn.

4. Emotional:
  - a. Has been in deep water for years.
  - b. Emotional stress due to his desire to learn and his efforts to learn, only to end in constant failure.
  - c. Misunderstood by principal, teachers, and classmates.

5. Home:
  - a. Mother is attractive and intelligent. She wishes to do everything possible for her son; has tried to help him with his lessons at home.
  - b. Home is a happy place. One younger brother.

C. *Remedial Program.*

1. Suggestions:
  - a. Correct vision immediately.
  - b. Meantime use material and method adapted to defective vision: Arrange for best possible light. Use a book with extra large print such as is used for sight-saving classes. Use chalk which makes a clear, distinct line. Write large script for him to read. Have him use a soft pencil which makes a heavy, dark line. Use a marker to help him keep the place.
  - c. Get expert medical advice on hearing as soon as he is adjusted to vision correction. Meantime adjust to his defective hearing: Try to be where he can see you when you speak to him. Speak very distinctly and in louder voice.
  - d. Ease the emotional upset by making it possible for him to succeed. Praise his efforts; help him to feel your interest; assure him that he will learn to read because you now understand why he has not learned and you know what to do about it. Inform other teachers and enlist their co-operation. Break down the rude, destructive attitude of the other children by your own example of friendly consideration and sympathy and by expecting the same of them.
  - e. Reinforce and stabilize the word images by kinaesthetic sense. Accomplish this by having him write his words in large script on the board or trace over words so written.
  - f. Use the usual drill procedures for words and phrases, but be sure of the size of the script or print.
  - g. Teach him phonograms and consonant sounds.
  - h. *Start on a first-grade level.* Use experience reading as well as a book and let him keep a record of progress.

2. Results:
  - a. A prescription was secured for vision, but glasses have not yet been provided because of an accident which has confined his mother to the house, and because of shortage of funds.
  - b. In spite of this, when taught as indicated, in a period of three months he has learned 150 new sight words, 20 phonograms, 26 consonant blends, and has been able to help himself considerably with new words.
  - c. On standard reading tests he has made a gain of one full year, his average having increased from 1.3 to 2.45.

- d. He is more composed.
- e. The diagnosis appears to have been entirely confirmed, and the urgent need for continuing the treatment established.

CASE NO. 37,[37] GIRL, UNCLASSIFIED, CA 13–7

A. *Nature of Difficulty.*

1. Emotional:
   - a. Believes herself to be defective.
   - b. Rebellious, refuses to return to school.
   - c. When brought to the examination under pressure, she refused to co-operate in the reading tests, saying, "So this is what they brought me here for! Well, you might just as well mark me zero and go on." An explanation of reading disabilities was then given to her, showing her how they may occur in bright children, and assuring her that she was not necessarily dumb just because she could not read.
   - d. After giving the Stanford-Binet Intelligence Test, the examiner said, "You see, Charlotte, these tests make me know that you are a bright girl. A dumb one could not pass them. You are only thirteen years old, yet you succeeded with some of the tests for fourteen- and sixteen-year-old people." She seemed pleased for a moment or two, as if wishing to be convinced, and then retorted: "But these tests don't count. They don't make any difference. It's whether or not you can read that tells how bright you are."
   - e. Co-operation was finally established, however, and the reading tests were completed.

2. Educational:
   - a. Grade level on Gray Oral Reading Paragraphs was 3.3. Excessive number of errors in consonants, reversals, addition of sounds, omission of sounds, repetition and addition of words.
   - b. Grade level in silent reading was 3.6 on Haggerty Sigma 1 Test.
   - c. Grade level on vocabulary tests was 3.6 on Iota Word Test and 3.9 on Monroe Word-Discrimination Test.
   - d. Grade level in other subjects: 3.1 on Ayres Spelling Scale and 5.3 on Stanford Arithmetic Computation Test.
   - e. She was a fluent mirror-reader and mirror-writer. She was right-handed, but preferred her left eye in sighting. She had great difficulty in forming visual-auditory associations, and in blending sounds in wordbuilding. She was never secure in recognition of complex word patterns, although she recognized the individual letters and words easily.

B. *Causative Factors.*

1. Intellectual:
   - a. Stanford-Binet MA of 14–3, with an IQ of 105. Trouble is not lack of general intelligence.

2. Emotional:
   - a. Pronounced emotional reaction toward school and reading in particular.
   - b. Resistant toward any attempt at further education.

---

[37] Adapted from Marion Monroe, *Children Who Cannot Read,* pages 167–170. Chicago: University of Chicago Press, 1932.

3. Educational:
   - a. A number of grade repetitions and unhappy experiences in school.
   - b. As teachers in the public school considered her defective in intelligence, she was placed in a private school for defective children. Here she remained for four years, until the date of the first tests.

C. *Remedial Program.*

1. Schedule:
   - a. Owing to her emotional resistance to school and her severe retardation, it was advised that she be taken out of school altogether for a period of nine months.
   - b. A trained tutor was employed for remedial work in reading. This was conducted on an individual basis daily for hour or hour-and-a-half periods. Although she proved to be a satisfactory pupil from the standpoint of attitude and effort, her extreme reading disability necessitated hours of patient, repetitive drill for each step of improvement obtained.
   - c. A second tutor was employed to work with her in arithmetic, history, geography, and content subjects.

2. Results:
   - a. Charlotte gained 4.0 years in reading achievement from September to June.
   - b. At the June examinations, all her scores were seventh- or eighth-grade level.
   - c. The following September she entered an eighth-grade class, from which she graduated creditably at the end of the school year.
   - d. She has since entered high school, where she is making a good adjustment. Her attitude toward her work has been excellent.
   - e. Qualitatively, Charlotte's reading still retains some aspects of her disability. For example, her reading has a slightly irregular speed quality characterized by periods of comparative facility and periods of blocking, in which the separate words must be studied. Hence she hesitates to read aloud before strangers. Her high-school work, however, requires little oral reading, and in silent reading her method of attack is not apparent.
   - f. On a repetition of the Binet-Simon Test fifteen months after the first test, her IQ had increased from 105 to 112, an improvement which the examiner attributed to improved reading ability and reduced emotional tension.

2. The General Technique of Diagnostic and Remedial Work in Reading

The foregoing case studies in reading are good illustrations of the various levels of educational diagnosis, with the exception of the first and the last. Some discussion is needed of the methods of locating pupils who require remedial attention, and of techniques for preventing difficulties in reading. A few generalizations regarding diagnostic and remedial procedures will supplement the case histories presented.

Locating retarded readers. Betts[38] suggests that pupils who have difficulty in learning to read may conveniently be divided into two classes: (1) Those who are below normal on the basis of general ability or intelligence, and (2) those who have specific learning disabilities in reading. While it is usually true that those who make low scores on tests of general intelligence will have difficulty in learning to read, it not infrequently happens that pupils of normal and even superior intelligence are retarded readers. The greatest amount of reading retardation is usually in the IQ range of 80 to 95. It is a well-known fact that a much larger number of boys than girls have reading difficulties. This is doubtless associated with the greater acceleration of language development in girls.

McCallister[39] proposes three methods of locating individuals requiring diagnostic and remedial treatment: (1) testing with standard tests, (2) collecting and analyzing pupils' cumulative records, and (3) analyzing school performance. Other methods used include informal teacher-made tests, observing pupils at work, interest inventories, actual trial with graded series of readers, or a study of eye movement during reading. But by far the most common method of locating pupils requiring remedial work in reading is to administer one or more standardized reading tests. Brief survey tests in reading, such as the Monroe, Thorndike-McCall, and Detroit, are often sufficient to reveal serious deficiencies. One advantage of such tests is that their use makes it possible to survey quickly the reading ability of a large number of pupils. A test of general intelligence, particularly one of the nonverbal type such as the Arthur Scale, is desirable. A general battery of achievement tests, such as the American School, Metropolitan, Modern School, Progressive, or Stanford, is a good means of locating poor readers. Any pupil whose reading ability falls behind his mental ability, his ability in other subjects, or his chronological age, is worthy of further study. Hildreth[40] makes the practical suggestion that any child of normal ability and experience who is a year retarded in reading at grades four, five, and six, or a year and a half retarded at grades six, seven, and eight, is in need of intensive study. A study[41] of 6,364 pupils representing the second to the sixth grades in eleven states found that about 15 per cent were retarded one or more years in reading.

---

[38] Emmett Albert Betts, *The Prevention and Correction of Reading Difficulties,* pages 1–2. Evanston, Illinois: Row, Peterson and Company, 1936.

[39] James M. McCallister, *Remedial and Corrective Instruction in Reading,* page 47. New York: D. Appleton-Century Company, 1936.

[40] Gertrude Hildreth, *op. cit.,* page 574.

[41] Clara L. Alden, Helen B. Sullivan, and Donald D. Durrell, "The Frequency of Special Reading Disabilities," *Education.* 62: 32–36, September, 1941.

Several writers [42] point out the desirability of supplementary test scores with other pertinent data. Perhaps most valuable of all are cumulative school records, which afford a complete developmental history of the individual as revealed by standard tests, teachers' marks, attendance, promotion, interests, and the like. That the judgment of a single teacher needs to be supplemented by actual testing is clearly indicated in a study by McCallister,[43] who found that approximately three fourths of the pupils referred to him by the classroom teacher as doing unsatisfactory work were really not retarded in reading, whereas several others not reported by the teachers were shown by the tests to be retarded readers capable of profiting from corrective instruction.

Analysis of reading difficulties. The principal value of survey reading tests is in locating pupils who are retarded in reading. Before appropriate remedial procedures can be devised, however, the specific nature of the difficulties must be discovered. This is the function of diagnostic tests. The Gray or Durrell tests will afford a detailed picture of the errors in oral reading. Among the well-known series of silent-reading tests which have considerable diagnostic value are those prepared by Durrell, Yates, Traxler, and Van Wagenen-Dvorak. The teacher in the elementary school will secure significant information regarding the more important reading skills from using such tests as the Iowa Silent Reading Test or the Iowa Every-Pupil Tests of Basic Skills. In the upper educational levels essential information is provided by such reading tests as the Co-operative, Schiammel-Groz or Van Wagenen.[44]

It is often a good idea to have the child read the same material orally as well as silently. As a rule, norms are comparatively unimportant in diagnosis. The way the pupil behaves in the reading situation is sometimes more revealing than the test score itself. Eye movements can be detected informally by the peephole method.[45] The material to be read is pasted on a cardboard with a small hole in the center, through which the teacher looks as the child reads. Excellent measurements of the visual functions employed in reading are afforded by the Keystone Ophthalmic Telebinocular, the Betts Ready to Read Tests, and the ophthalmograph. Although the evidence, on the whole, has been favorable to these techniques, contradictory results have been reported.[46] At best eye movements are merely symptomatic of reading difficulty rather than causal factors. As a rule, when appropriate reading experience is provided, eye movements will take care of themselves.

---

[42] Paul Witty and David Kopel, *Reading and the Educative Process,* Chapter III. Boston: Ginn and Company, 1939.

[43] James M. McCallister, *op. cit.,* page 63.

[44] For a useful description of these and other standard tests of silent reading, see: Constance M. McCullough, Ruth M. Strang, and Arthur E. Traxler, *Problems in the Improvement of Reading,* pages 125–133. New York: McGraw-Hill Book Company, 1946.

[45] W. R. Miles and David Segel, "Clinical Observation of Eye Movements in the Rating of Reading Ability," *Journal of Educational Psychology,* 20: 520–529, October, 1929.

McCallister [47] has made an extensive study of reading deficiencies found in pupils in the University of Chicago High School. Table 34 summarizes the results for a group of eighteen pupils. McCallister [48] also made a careful study of reading difficulties in American history, mathematics, and general science. He visited the classes, analyzed the reading activities presented by the teaching methods and materials, and determined the resulting reading difficulties by a careful analysis of pupils' written reports and by observation of their work habits. The fifty difficulties located were summarized under six headings as follows: Faulty methods of attack, inability to recognize relations, lack of knowledge of subject matter, deficiencies in vocabulary, inaccuracies, and lack of clearness in directions given to pupils. This study illustrates fruitful possibilities open to any teacher. It is perhaps surprising to discover that the difficulties in mathematics are more numerous than those in American history, as well as wholly different in character.

Locating the causative factors in reading disability. Typically, diagnosis proceeds from symptoms to underlying causes. High temperature and pain are useful danger signals to the doctor, but he recognizes that it is useless to treat symptoms. Curative measures must be based on the factors which are causing the high temperature and pain. Exactly the same relationship exists in educational maladjustment. It is often extremely difficult to determine accurately the basic causative factors. The teacher must resist the tendency to offer too easy an explanation, such as low mentality or organic deficiency. Most cases of reading disability,[49] for example, are due to no single, isolated cause, but to a number of interrelated factors. Many factors may be either cause or effect. Poor school attendance, truancy, withdrawal behavior, and dislike for reading are examples of factors that undoubtedly hinder the pupil's present

---

[46] Henry A. Imus, John W. M. Rothney, and Robert M. Bear, *An Evaluation of Visual Factors in Reading,* 144 pages. Hanover, N. H.: Dartmouth College Publications, 1938.

[47] James M. McCallister, "Character and Causes of Retardation in Reading among Pupils of the Seventh and Eighth Grades," *Elementary School Journal,* 31: 35–43, September, 1930.

[48] James M. McCallister, "Reading Difficulties in Studying Content Subjects," *Elementary School Journal,* 31: 191–201, November, 1930.

[49] Mary L. Preston, "The School Looks at the Nonreader," *Elementary School Journal,* 40: 450–458, February, 1940.

Table 34

READING DEFICIENCIES FOUND AMONG EIGHTEEN PUPILS IN THE SEVENTH AND EIGHTH GRADES AND THE FREQUENCY OF OCCURRENCE OF EACH (AFTER McCALLISTER)

Deficiency	Pupil 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Frequency
Deficiencies in comprehension and interpretation:																			
Reading ability not developed to the point that pupil could interpret with facility materials of the level of difficulty found in regular textbooks	--	--	--	--	x	x	x	--	x	--	--	x	x	x	x	--	x	x	10
Inaccuracy in interpretation	x	x	--	--	x	--	x	x	--	--	x	--	--	--	x	--	x	x	9
Excessive re-reading required for interpretation	x	--	x	--	--	x	--	--	x	--	--	x	x	x	x	--	--	x	9
Word-reading with little attention to content	--	x	--	--	x	--	--	--	x	--	--	--	--	--	--	--	x	--	4
Rapid but superficial reading	--	--	--	--	--	--	--	x	--	--	x	--	--	--	--	--	--	--	2
Inability to answer thought-provoking questions based on reading materials	--	--	--	--	--	--	--	--	--	--	--	x	--	--	--	--	x	--	2
Inability to formulate conclusions on basis of passages read	x	x	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	2
Deficiencies in rate of reading:																			
Slow rate of silent reading	x	x	--	x	x	x	x	--	x	x	--	x	x	x	x	x	--	x	14
Slow rate of oral reading	x	--	--	--	x	x	x	--	x	--	--	--	x	--	x	--	--	x	8
Deficiencies in fundamental reading habits:																			
Numerous regression movements	x	x	x	x	x	x	x	--	x	x	--	x	x	--	x	x	--	x	14
Narrow span of recognition	x	x	x	x	x	x	x	--	--	--	--	x	x	x	--	x	x	x	13
Inaccurate return sweep of the eye	x	--	--	x	x	x	x	--	x	x	--	--	x	x	--	--	--	x	10
Irregular rhythm in silent reading	x	--	--	x	x	x	x	--	--	x	--	--	--	--	x	--	--	x	8
Inaccuracies in recognition of familiar words	--	--	--	--	--	x	x	x	x	--	--	--	--	--	x	--	--	x	6
Excessive vocalization	--	--	--	x	--	x	--	--	--	x	--	--	--	--	--	x	x	--	5
Frequent moments of confusion	--	x	--	--	--	x	x	--	x	--	--	--	--	--	--	--	--	--	4
Inability to cope with new words	--	--	x	--	x	--	--	--	--	--	--	--	--	--	--	--	x	--	3
Oral reading jerky and expressionless	--	--	--	--	--	x	x	--	--	--	--	--	--	--	--	--	--	--	2
Forward movements of the eye too long	--	--	--	--	--	--	--	x	x	--	--	--	--	--	--	--	--	--	2
Excessive head movements	--	--	--	--	--	--	--	--	--	x	--	--	--	--	--	--	--	--	1

performance, but they may also be the effects of many poor performances in the past. A common defect of diagnosis is to neglect the individual's past history, especially his experience in school.

Hildreth characterizes the poor reader [50] as follows:

> The deficient reader is more apt to be a boy than a girl; he is apt to be deficient in some phase of language usage, or to come from a foreign home. He is apt to be the child who missed considerable school work in the first grade; to be a restless, unstable child, difficult to control; or a babyish individual, immature in emotional development. For one of his age, he is apt to be below average in mental ability, and to have difficulty in giving attention and in concentrating. He lacks experiences common to most children, and is more apt to have sensory and physical deficiencies.

Many large school systems have established reading clinics at which the severest cases may be expertly diagnosed. The techniques employed in these schools, although more elaborate than are necessary or practical for the ordinary classroom teacher, are worthy of careful study for the light they shed upon the problem. The operation of the Pittsburgh clinic is described as follows: [51]

> Four examiners meet every child who is referred to the clinic. The medical doctor gives the child a complete physical examination, including X-ray, basal metabolism, and any other laboratory tests needed to make a complete study of the child's physical condition. A psychologist gives each child a series of intelligence tests. Usually the Stanford-Binet examination is supplemented by the Grace Arthur Performance Scale in order to obtain ratings on tests of different types of intelligence. Another examiner gives each child a series of psychological laboratory tests including tests of visual acuity, eye-muscle balance, tests of hearing at various ranges of pitches with an audiometer, and photographs of eye movements during reading. Still another examiner gives the child a series of achievement tests and tests for various types of reading skills. Parents are interviewed to obtain the child's personal history and to determine any factors in the home environment which may be contributing to his inability to read. The child's school history is obtained and studied for those factors which may help to explain his reading failure.

In general, at the present time there is less tendency than formerly to ascribe deficiencies in reading to specific mental disabilities, such as "word blindness" or "mixed dominance," and a greater tendency to regard most visual and auditory defects and limitations of character and personality as contributory rather than as direct causal factors. On the contrary, greater emphasis is placed on environmental factors more directly within the control of the teacher, such as inappropriate instructional materials and methods.[52]

---

[50] Gertrude Hildreth, *op. cit.,* pages 372–373.

[51] Marion Monroe, "Diagnostic and Remedial Procedures in Reading," *The Educational Record,* 19: 107–108, Supplement for January, 1938.

[52] For an excellent recent discussion of causative factors in reading see: Paul Witty and David Kopel, *Reading and the Educative Process,* Chapters VII–IX. Boston: Ginn and Company, 1939.

A recent book points out that although "there is no one best program" of remedial reading any successful program is likely to have the following basic features: [53]

1. Select students for special reading classes who are in the lowest fourth of their class on a standardized reading test, have given other evidence of reading difficulty, and want to improve their reading efficiency.

2. Provide time for the class in the student's regular program—preferably, two class periods a week and an additional conference period, for individual help.

3. Give students as much responsibility as possible for planning and carrying out a realistic program, suitable to their abilities and their present and future needs for reading.

4. Administer additional silent and oral reading tests, both standardized and informal, for further diagnostic information.

5. Create an atmosphere of optimism reinforced by experiences of success in reading.

6. Use on-going activities and already established interests as natural incentives to read.

7. In the beginning, supply reading materials of intrinsic interest that is at, or slightly below, the student's present reading ability.

8. Use assignments in other subjects as practice material and as a basis for specific instruction.

9. Give drill whenever necessary in individual cases or in the group as a whole to overcome specific reading difficulties.

10. Help each student to keep a record of his progress in reading.

11. Help students to make the transition from the reading class to reading in other classes and to voluntary reading.

12. Evaluate changes in reading ability made during the special class.

It will be noted that the program outlined indicates clearly the close relation between testing and remedial instruction. Both standardized and non-standardized tests are employed. Any alert teacher can prepare diagnostic and instructional materials that will compare favorably with most of those commercially available today. Wrightstone,[54] for example, describes a most effective program of diagnostic and remedial work carried out by the regular teaching staff:

The typical class project in remedial reading which is reported in this article demonstrates that it is not necessary for the teacher to be a specialist in the techniques of diagnostic and remedial reading. A few diagnostic instruments—most of them teacher-made—can be employed to aid the teacher. The application of some common sense remedial measures to overcome the lack of adequate reading skills and abilities is almost always accompanied with favorable results.

---

[53] Constance M. McCullough, Ruth M. Strang, and Arthur E. Traxler, *op. cit.*, pages 224–225. Reprinted by permission of the publishers, McGraw-Hill Book Company, Inc.

[54] J. Wayne Wrightstone, "Diagnostic Reading Skills and Abilities in the Elementary Schools," *Educational Method*, 16: 248–254, February, 1937.

The case studies reported illustrate most of these principles and demonstrate their effectiveness in teaching reading. They appear, however, to be equally successful when applied to other subjects. One of the most difficult problems is usually the correction of the pupil's attitude toward himself, for by the time he has gone through three or four grades without learning to read, he is quite fully convinced that he is so stupid that he can never do so. A key problem is the improvement of morale. Blair[55] recognizes only two absolutely essential conditions for the improvement in reading on the high-school level: a desire to read and an opportunity to read abundant materials of suitable character.

Preventive diagnosis in reading. But important as is remedial work in reading, the prevention of reading difficulties is far more important.[56] Most authorities in reading agree that a child can learn to read unless he is hopelessly deficient mentally. The fact that he often does not do so, or does so only after many failures, is all the more tragic because his failure need not happen. Poor readers are made and not born. As Betts says, *"Poor teaching, in a larger sense, is the chief cause of reading retardation."*[57] A fuller statement by the same author follows:[58]

> The point of view of the writer is that most reading difficulties could be prevented by increasing the entrance age for the first grade, by revising the first-grade program of instruction, by providing a greater quantity of primary-reading material with varied content, by grouping children in terms of general readiness for a given reading program, by beginning instruction with the learner's interest, and by the correction of physical defects. The emphasis in the reading program should be on prevention rather than correction.

Altogether too few teachers appear to realize that the transition from home to school is a crisis in the life of any child. When he crosses the threshold of the school, he sets his feet in a strange, new world. Here for the first time he leaves his familiar environment and must learn to make his way in a new one. He must accept directions from a total stranger, the teacher, instead of from his mother. Instead of the neighbors' children with whom he is accustomed to play, he is confronted with a roomful of children whom he has never seen before. In place of his beloved toys he is given strange materials with which to work. Moreover, as if adjustment to this world of strange people and things were not enough, the first grade presents a wholly new world of abstract ideas and symbols. No wonder he often feels bewildered and helpless.

---

[55] Glenn Myers Blair, *op. cit.,* page 169.

[56] Paul Witty and David Kopel, *op. cit.,* Chapter VI.

[57] Emmett Albert Betts, *op. cit.,* page 52.

[58] *Ibid.,* page 9.

And not without cause. The plain facts are usually that he is indeed entering an inhospitable world of the fang and the claw, where only the "fittest" survive. The records show that many are not "fit." The largest number of school failures is usually in the first grade, and more often in reading than in anything else. Fortunately an increasing number of school people are recognizing that it should be somebody's business to see that the child is fit, as evidenced by the various tests devised to determine reading readiness.

Several years ago a distinguished psychologist emphasized the wisdom of this motto for teachers: "Don't cut off the tadpole's tail." Man can speed up the process somewhat by arranging favorable conditions for growth, but in the final analysis the only way to get a frog is to wait for him to develop. Much the same thing is true of teaching. The child will learn to read, as he learns to talk or to walk, when he is ready to do so, and not before. It is possible, however, to facilitate the development of readiness by skillful teaching. Efforts to force the learning process before the stage of readiness are usually ineffective, often tragic.

Gray says that there is general agreement today "that successful reading at all grade levels is conditioned in large measure by the physical, mental, emotional, and social maturity of the learners and by proper adaptation of instruction to their needs." [59] He enumerates seven "essential prerequisites to reading," as follows: [60]

1. Wide experience as a background for interpretation.
2. Reasonable facility in the use of ideas.
3. Reasonable command of simple English sentences.
4. A relatively wide speaking vocabulary.
5. Accuracy in enunciation and pronunciation.
6. Reasonable accuracy in visual and auditory discrimination.
7. Keen interest in learning to read.

It is well to note that the statements above are explicit as to the types of readiness, but that they are somewhat indefinite as to the amount. Just how extensive is a "wide experience" or a "relatively wide speaking vocabulary"? It is doubtless impossible to make an exact quantitative statement that will fit all situations. Much depends upon the child and the type of reading program offered, as well as upon the experience and skill of the teacher. The published norms on general intelligence tests and on specific aptitude tests for reading, called reading-readiness tests, are usually less important than are local norms derived from experience in the particular situation. A minimum mental age of six years, or of six and one half years, is often suggested, but there is evidence to show that certain types of children can handle successfully certain types of reading programs much below this level.

---

[59] William S. Gray, "The Nature and Organization of Basic Instruction in Reading," *Thirty-Sixth Yearbook of the National Society for the Study of Education, Part I,* page 79. Bloomington, Illinois: Public School Publishing Company, 1937. (Quoted by permission of the Society.)

[60] *Ibid.*, pages 82–84.

Evidence available suggests that the total score on readiness tests is less important than the diagnostic value of the scores on the separate parts of the tests. There seems no good reason why readiness tests cannot be developed for determining when the child should take up the various phases and levels of reading and other school subjects, as well as when to begin the study of these subjects. Furthermore, readiness should not be regarded as some mysterious entity whose existence is wholly due to inner growth. On the contrary, readiness in all subjects is at least in part the product of experience largely within the control of the school.

Research is needed to determine at what age learning to read is educationally and socially most fruitful to the child, and then to determine the most appropriate materials and methods for this age. It is much sounder educational practice to adjust the school to the child than it is to attempt to adjust the child to the school. The diagnostic program in reading and in all other subjects may properly have as its immediate objective the correction of difficulties that now exist, but its ultimate goal should always be the prevention of similar occurrences in the future.

Concluding statement. Diagnosis is the most important function of measurement in any subject and on any educational level. Its immediate purpose is to point out where remedial measures must be applied to correct existing deficiencies, but its ultimate goal is to prevent the recurrence of similar weaknesses in the future. The general principles have been set forth and illustrated by the subject of reading. That they apply equally well to other subjects can be readily seen by consulting the references cited in this chapter.

SELECTED REFERENCES FOR FURTHER READING

Betts, Emmett Albert, *Foundations of Reading Instruction with Emphasis on Differential Guidance.* New York: American Book Company, 1946, 758 pages.

Betts, Emmett Albert, *The Prevention and Correction of Reading Difficulties.* Evanston, Illinois: Row, Peterson and Company, 1936. 402 pages.

Blair, Glenn Myers, *Diagnostic and Remedial Teaching in Secondary Schools.* New York: The Macmillan Company, 1946. 422 pages.

Brueckner, Leo J., and Melby, Ernest O., *Diagnostic and Remedial Teaching.* Boston: Houghton Mifflin Company, 1931. 598 pages.

Brueckner, Leo J. and others, "Educational Diagnosis," *Thirty-Fourth Yearbook of the National Society for the Study of Education.* Bloomington, Illinois: Public School Publishing Company, 1935. 523 pages.

Durrell, Donald D., *Improvement of Basic Reading Abilities.* Yonkers: World Book Company, 1940. Chapters II, XIII, and XIV.

Fernald, Grace M., *Remedial Techniques in Basic School Subjects.* New York: McGraw-Hill Book Company, 1943. 349 pages.

Greene, Harry A., Jorgensen, Albert N., and Gerberich, J. Raymond, *Measurement and Evaluation in the Elementary School.* New York: Longmans, Green, & Company, 1942. Chapters XIII and XVIII.

Hildreth, Gertrude, *Learning the Three R's,* Second Edition. Philadelphia: Educational Publishers, Inc., 1947. 897 pages.

Kelley, Truman Lee, *Interpretation of Educational Measurements.* Yonkers: World Book Company, 1927. Chapters IV, V, and VI.

Lee, J. Murray, *A Guide to Measurement in Secondary Schools.* New York: D. Appleton-Century Company, 1936, Chapters VIII and IX.

McCall, William A., *Measurement.* New York: The Macmillan Company, 1939. Chapter XXI.

McCullough, Constance M., Strang, Ruth M., and Traxler, Arthur E., *Problems in the Improvement of Reading.* New York: McGraw-Hill Book Company, 1946. 406 pages.

Orleans, Jacob S., *Measurement in Education.* New York: Thomas Nelson and Sons, 1937. Chapters X and XI.

Symonds, Percival M., *Diagnosing Personality and Conduct.* New York: D. Appleton-Century Company, 1931. 602 pages.

Tiegs, Ernest W., *Tests and Measurements in the Improvement of Learning.* Boston: Houghton Mifflin Company, 1939. Part II.

Traxler, Arthur E., *The Use of Test Results in Diagnosis and Instruction in the Tool Subjects.* New York: Educational Records Bureau, 1942. 80 pages.

Witty, Paul, and Kopel, David, *Reading and the Educative Process.* Boston: Ginn and Company, 1939. 374 pages.

CHAPTER XIV

School Marks

A. The Problem of Marks

Present status of marks. In recent years several surveys have been made of marks and marking systems. All of these show that there exists a wide diversity both in theory and in practice. Billett[1] studied the best practice existing in American secondary schools in 1932. Counting minor variations, he found 100 different marking systems in use; of these, two thirds had been changed within ten years. Bixler summarized the literature on marking for the years 1933 to 1936 and concluded that there was "general dissatisfaction with the present marking system, but, as yet, little agreement as to the direction in which to go."[2] Three years later he began a similar review of the intervening period with the statement, "Dissatisfaction with marking systems has increased."[3] Variations exist not only among school systems and among schools in the same system, but also within the same school. For example, Billett found that only half of his superior schools used the same marking plan in both the elementary and the secondary school.[4] On one important point there seems to be fair agreement: "Among schools of all levels a system of five symbols, four of which are passing and one failing, is by far the most common."[5]

Even in the same elementary or high school the widest possible disagreement may exist. This disagreement occurs not only in standards used and in factors considered in assigning marks, but also in the significance attached to them. Some teachers appear to hold that passing examinations and accumulating credits and grades is the acme of all academic achievement, the pupils' chief aim and end in life. In the same building or department other teachers may hold that tests and marks of any kind, if not creations of Satan, are, at any rate, relics of barbarism whose existence is more appropriate in a museum than in a modern school and whose elimination would bring, immediately and automatically, the educational millennium.

---

[1] Roy O. Billett, *Provisions for Individual Differences, Marking and Promotion,* National Survey of Secondary Education, Monograph 13, pages 424–472. Washington, D. C.: United States Office of Education, 1932.

[2] Harold H. Bixler, "School Marks," *Review of Educational Research,* 6: 171, April, 1936.

[3] Harold H. Bixler, "School Marks," *Review of Educational Research,* 9: 172, April, 1939.

[4] Roy O. Billett, *op. cit.*, page 427.

[5] C. W. Odell, "Marks and Marking Systems," in *Encyclopedia of Educational Research,* edited by Walter S. Monroe, page 698. New York: The Macmillan Company, 1941.

In spite of this widespread dissatisfaction with school marks and considerable evidence of a reduced emphasis upon them in the elementary school, it seems most unlikely that marking systems are going to be abandoned by American schools in the near future. For example, a survey [6] of 35 school districts in the progressive state of California, representing all grades, revealed that more than 70 per cent of the teachers and principals, and more than 80 per cent of the pupils and parents voted in favor of school marks and report cards. And yet Bixler tells us that, up to the present time, "no one has studied the fundamental problem of whether or not marks are necessary in any comprehensive way." [7] It appears that, notwithstanding the acknowledged limitations of marks, no satisfactory substitute has yet been found.[8] Any discussion of measurement must consider the problem of school marks. Odell [9] has called attention to "many unanswered questions and unsolved problems in this area" on which further research is needed.

Need for a marking system. Concerning one phase of the marking problem there is rather general agreement. It is recognized that for marks to have meaning there must be a greater degree of uniformity than now exists. But before there can be consistency in practice, there must be agreement in theory. The first need, therefore, is for an intelligent marking policy, a plan of action that will insure a reasonable degree of uniformity. The second need is for a sound marking technique for putting the plan into operation. Lamson [10] makes a strong plea for what is termed "a philosophy of marks."

B. Essentials of a Satisfactory Marking Policy

Importance of a group policy. The first requirement of a satisfactory marking policy is that it be a group policy, arrived at after discussion and deliberation by all the members of the teaching and administrative staffs. It should not be formulated or "borrowed" by the administrative organization and then "handed down" to the classroom teachers. The faculty must understand it and believe in it. The best way to bring this situation about is to allow the faculty a voice in making the policy. It must be the result of co-operative effort. The responsibility of the administrative staff is to educate the faculty and the public in order to bring about an intelligent understanding and appreciation of the marking problem. The educational value of *making* the policy should not be overlooked.

---

[6] Ernest W. Tiegs, *Tests and Measurements in the Improvement of Learning,* pages 416–417. Boston: Houghton Mifflin Company, 1939.

[7] *Review of Educational Research,* 6: 173, April, 1936.

[8] *Cf.* Ivan H. Linder, "Is There a Substitute for Teachers' Grades?", *American School Board Journal,* 101: 25–26, July, 1940.

[9] C. W. Odell, *op. cit.,* page 701.

[10] Edna E. Lamson, "The Problem of Adequate Evaluation of the College Student's Achievement," *Educational Administration and Supervision,* 26: 493–507, October, 1940.

Administrators often take too much for granted. The situation in the elementary and secondary school is frequently little better than that in college, which has been described as follows: [11]

> Persons in charge do not realize that there is any problem involved. Everybody gives grades; everybody must know how to give them. It is like reading or writing. To attempt to make any suggestions would be an affront.

The original formulation of the marking policy should be the subject of a series of teachers' meetings. The major points to be considered are suggested by the topic headings in this chapter. From time to time after the original formulation, the policy should come up for review and possible revision. Any significant developments should be brought to the attention of the staff. All new teachers should be carefully instructed regarding the policy. Billett reports that four fifths of the superior secondary schools studied hold individual or group conferences on methods of marking.[12] Whenever any material changes are made, they should be fully explained both to the pupils and to their parents. One of the best ways of doing this is to use school marks as a topic for discussion and debate among the pupils, possibly in English classes or in the social studies. In recent years there has appeared a trend toward permitting students a larger voice in all aspects of evaluation.

Function of marks. Why give marks, anyway? This is logically the question with which to begin any discussion of marks and marking systems. If there appear to the teaching staff and school administrators to be no good reasons, or insufficient reasons, for giving marks, the problem is not how to secure better marks but how to find satisfactory substitutes. On the other hand, if there appear to be valid reasons for marks under existing conditions, the problem then is how to devise a marking policy that will serve these functions to the maximum degree.

---

[11] Ralph B. Spence, *The Improvement of College Marking Systems,* page 2. New York: Bureau of Publications, Teachers College, Columbia University, 1927.

[12] Roy O. Billett, *op. cit.,* pages 438–439.

It is important at the outset that everyone connected with the school understand as clearly as possible the purposes that marks are intended to serve. Marks are always means to ends, never ends in themselves. It will be recalled that all the purposes measurement is expected to serve have been grouped under two headings, administrative and instructional. These may then be conveniently subdivided, as follows:

I. Administrative functions:
    1. Classification and promotion.
    2. Guidance.
    3. Evaluation.
    4. Public relations.

II. Instructional functions:
    1. Motivation.
    2. Practice or drill.
    3. Diagnosis.
    4. School marks.

It is well to recognize a distinction between the nature and the function of tests and of marks. A teacher may believe strongly in the educational value of tests and not believe in school marks at all. It is undoubtedly easier to make a case for tests than for marks. Suitable tests are among the most important instruments of diagnosis available, while school marks have little or no diagnostic value. Tests may serve a practice or drill function, but marks do not. It can be seen that the instructional value of marks depends upon how well they motivate learning. The evidence for and against this claim has already been considered in Chapter XI. A study by Coble[13] indicates that high-school pupils were motivated less by *Pass* and *Failure* than by the usual marks, *A, B, C, D, E*. Even if evidence is available to show that pupils are moved to greater effort because of the fear of failure or the hope of reward in the form of a mark, the school must still carefully consider whether such extraneous motivation is consistent with its educational philosophy. It does not necessarily follow that because a pupil is genuinely interested in learning for its own sake he will be uninterested in knowing about his progress. Even the most enthusiastic golfer would hardly neglect to keep his score or be content with merely *Win* or *Lose*. Steiner[14] reports that on secret ballots 80 per cent of high-school pupils voted consistently against changing to *Pass* and *Fail* from their present more discriminating six-letter system. Williams[15] suggests that at the present time neither parents nor higher institutions of learning are ready for such a change.

---

[13] Robert Coble, Master's Thesis, Pennsylvania State College, 1936, summarized in *The Pennsylvania State College Studies in Education,* No. 19, pages 17–18, 1937.

[14] M. A. Steiner, "Democracy vs. School Regimentation," *School and Society,* 53: 106–110, January 25, 1941.

The administrative claims of school marks rest on a more secure foundation, however. While marks are *given* by the teachers, they are *used* mainly by the administrators. Marks undoubtedly *do* serve the four important administrative functions enumerated above, and in some form have doubtless been doing so since schools began. Table 35, taken from Billett,[16] shows that marks are used for all these purposes by the majority of the 258 superior secondary schools studied.

TABLE 35

PURPOSES SERVED BY MARKS IN 258 SUPERIOR SECONDARY SCHOOLS (AFTER BILLETT)

PURPOSE	FREQUENCY	
	Number	*Per Cent*
1. Keeping parents informed of pupil's progress	244	95
2. Furnishing a basis for promotion	238	92
3. Furnishing a basis for graduation	212	82
4. Motivating pupils	194	75
5. Furnishing a basis for the awarding of honors	190	74
6. Furnishing a basis for guidance in the election of subjects	158	61
7. Furnishing a basis for guidance in college recommendation	155	60
8. Furnishing a basis for determining extent of participation in extracurricular activities	133	52
9. Furnishing a basis for guidance in recommendation for employment	113	44
10. Furnishing a basis for awarding credit for quality	100	39
11. Furnishing a basis for research	50	19

How well they serve these functions is a different question. Certainly the functions themselves are important. Some basis for classifying and promoting pupils must be used, for example, but if chronological age or standard tests can do the job better, it would be difficult to justify the use of school marks for this purpose. That school marks afford useful data in guidance has been amply demonstrated. It has been pointed out that, to the counselor, marks are "equivalent to the inspection of the patient's tongue by the doctor: something may look wrong but it will take more refined diagnoses to locate the malfunction."[17] It is worth noting that, if frequency of mention is a measure of importance, the schools in Table 35 regard the public relations function as the most important of all, since "keeping parents informed of pupil's progress" heads the list, with 95 per cent. As commonly used, however, this information is highly illusory. A personal conference with the parent, or a personal note regarding the pupil's status in relation to the various objectives of instruction, is likely to be far more truly informative, and these alternatives have made a considerable amount of headway in the elementary school.[18] The comparative merits of these methods are worthy of careful consideration.

---

[15] L. A. Williams, *Secondary Schools for American Youth,* page 323. New York: American Book Company, 1944.

[16] Roy O. Billett, *op. cit.,* page 449.

The extensive literature on the subject should be examined in formulating the marking policy of the school or school system. Here, as always, objective evidence is more important than personal opinion. The alleged limitations of marks, as well as the claims made for them, should be critically considered.

Sources of weakness in marks. The more clearly teachers understand the sources of weakness in existing marks, the more likely they are to see what should be done to strengthen them. Just why are teachers' marks often considered to be lacking in both validity and reliability?

The principal reason is that teachers often allow various extraneous factors to enter into the determination of the marks. Even if the faculty are blissfully ignorant of what is happening, the pupils appear to recognize the situation well enough, as is shown in the following jingles:

If
(With apologies to Kipling)

If you can fool your prof without his knowing
That you are shooting him a line of bull,
And, while this bull from you is freely flowing,
By degrees you get with him a pull;
If you can force yourself by mighty effort
To laugh when he dispenses his stale jokes

---

[17] E. G. Williamson and J. G. Darley, *Student Personnel Work,* page 123. New York: McGraw-Hill Book Company, Inc., 1937.

[18] J. W. Poynter, "Blind Man's Buff," *School Executive,* 59: 13–14, September, 1939.

And, when all your fellow students miss a question,
You read the answer calmly from your notes;
If you can fill the unforgiving hour
With sixty minutes worth of spoofing done,
Yours is the course, and three hours credit with it,
And, what is more, you'll make an A, my son.

At least one pedagogue has recorded in rhyme his reflections on the marking problem and has given audible expression to the doubts which have perplexed many others. The verse bears the appropriate and revealing title, *Mystery.*

Mystery

By R. J. Bretnall[19]

Bear with me while I here relate
How I am forced to meditate,
Most actively to cogitate
On pupils' academic fate.
When grading periods roll around,
I sit me down with thoughts profound
To render judgments somewhat sound.

We know our school is quite progressive
With eye alert for traits recessive
And personality possessive.
Yet, sure as downward runs the water,
Each mother's son and father's daughter
Is held for high scholastic slaughter.
What mark for James? I cannot tell:
In oral English he does well. . . .

Alack, alack, he cannot spell.
He knows the country of the lama,
Excels in geographic drama
But cannot locate Atacama;
Though tales of statesmen he'll relate,
Refuses, straight, to learn the date
When Idaho became a state.
What shall I give, now let me see—
An A, a B, a C, or D,
An E, or F, what shall it be?

An A, we certainly agree,
Will fill his soul with ecstasy;
His boyish heart will throb in glee
Should I award to him a B;
No trouble would arise from C;
But if I dare put down a D
Parental wrath descends on me.

---

[19] *The Clearing House,* 11: 227, December, 1936.

I search with thoughtful, deep intent
For something that will represent
A guess, perhaps intelligent.
My brain reels on in wild congestion;
Down goes a mark at some suggestion—
And then I ask myself this question,
"Is this his grade, or my digestion?"

Teachers often discover an embarrassingly close relationship between their own "popularity" and the marks they are accustomed to "give." There seems indeed to be more truth than poetry in the following lines:

Some teachers give high marks while others give low;
Most students like high marks, so where do they go?
To teachers whose high marks are easy to get,
While hard-working teachers have seats to let.

Gillis[20] found a definite tendency in the freshman year of college for girls to receive higher marks in classes taught by men than in classes taught by women, and for boys to receive higher marks in classes taught by women than in classes taught by men. Studies on the high-school level, however, have shown that "men teachers favor boys and women teachers favor girls in the awarding of marks."[21] Since the majority of college teachers are men and the majority of high-school teachers are women, there is great probability that on all academic levels girls will receive higher average marks than boys whose ability and achievement are equal, when judged by norms on standard tests.[22] It seems too bad that the marks received by certain individuals are conditioned more by the contours of the face than by the contents of the head. Other studies have shown that the pupil's handwriting, conduct, language ability, seating position in the class, and ratings on such personality traits as respect for authority and co-operativeness are significant factors in determining his mark, as well as the condition of fatigue or boredom the teacher happens to be in when it is awarded.

Gillis[23] obtained the opinions of 1,000 college teachers, equally divided between the North-Central and the Southern Associations of Colleges and Secondary Schools, as to the factors considered in determining a mark. The results for mathematics and English, summarized in Table 36, show an extremely wide diversity of practice in regard to both the factors considered and the weight allowed to each. The situation appears to be equally bad in both subjects and in both associations. It is probably no better in other geographical areas or on other educational levels. Harris [24] has prepared a useful summary of 328 studies.

---

[20] Ezra L. Gillis, "Marking in Higher Institutions," *Proceedings of Kentucky Colleges and Universities,* 2: 93–108, 1932.

[21] Roy O. Billett, *op. cit.,* page 428; Clifford Swenson, "The Girls are Teachers' Pets," *Clearing House,* 17: 537–540, May, 1943.

[22] Giles M. Ruch and David Segel, *Minimum Essentials of the Individual Inventory in Guidance,* pages 21–22. Washington, D. C.: United States Office of Education 1939.

[23] Ezra L. Gillis, *op. cit.,* pages 99–103.

Proper bases for marks. What, then, are the proper bases for school marks? It is probably well to break this into two questions: What factors should be considered? What proportional weight should each have?

The answer to the first question must be: *In determining any mark, only those factors should be taken into account which afford evidence of the degree to which the pupil has attained the objectives set up for that particular course.* Daily work, class tests, oral quizzes, and final examinations are examples of factors which indicate progress toward the objectives of instruction. In other words, they may be accepted as evidences of scholarship. Attendance, effort, attitude, conduct, and the like, on the other hand, should receive no direct consideration, since indirectly and automatically they affect the pupil's test scores anyway. For example, if a pupil has been absent from class a week, it is not necessary to make an arbitrary deduction from the test scores actually made, since, presumably, if class attendance is really important, the scores have already been reduced by the amount he would have learned had he been present. Any arbitrary deduction for such factors is not only unnecessary but unfair. In like manner, except in English classes, no deduction for deficiencies in language usage, spelling, and handwriting should be made, unless such deficiencies obscure the meaning.[25]

Many teachers will object to leaving out of consideration such factors as the pupils' attitude, effort, conduct in the class, and various personality traits. These are certainly important and should be taken into account in some way. But if only one mark is given, it should be a mark in *scholarship,* and not a hodgepodge of miscellaneous items. A second mark, in *citizenship,* including most of the above items, has been added by many schools with a considerable amount of success. It is better still, no doubt, to provide a separate mark or rating to indicate growth in each characteristic or trait deemed important by the school. Even if many of these traits are highly subjective, the average judgment of the group should have considerable validity. It has also been found helpful to have the

---

[24] Daniel Harris, "Factors Affecting College Grades: A Review of the Literature, 1930–1937," *Psychological Bulletin,* 37: 125–166, March, 1940.

[25] For a fuller discussion of this point, see Chapter VI.

TABLE 36

FACTORS CONSIDERED AND THE WEIGHTS ALLOWED TO EACH BY TEACHERS OF MATHEMATICS AND ENGLISH IN THE NORTH-CENTRAL AND SOUTHERN ASSOCIATIONS OF COLLEGES AND SECONDARY SCHOOLS (AFTER GILLIS)

FACTOR	MATHEMATICS				ENGLISH			
	NORTH-CENTRAL		SOUTHERN		NORTH-CENTRAL		SOUTHERN	
	Range	*Mean*	*Range*	*Mean*	*Range*	*Mean*	*Range*	*Mean*
1. Attainment in subject	0–100	64.9	25–100	63.9	0–100	66.8	0–90	55.1
2. Attitude toward instructor	0–45	.9	0–10	1.7	0–10	1.3	0–1	1.0
3. Attitude toward subject	0–25	2.7	0–10	2.7	0–25	3.0	0–10	2.7
4. Attendance	0–10	1.9	0–15	5.6	0–33	2.3	0–15	3.9
5. Industry and application	0–40	3.2	0–10	4.6	0–15	2.0	0–5	1.4
6. Time expended by student	0–15	1.4	0–10	2.3	0–10	.7	0–5	.8
7. Originality and initiative	0–50	10.0	0–20	7.3	0–40	9.5	0–40	12.3
8. Deportment	0–10	.6	0–5	1.4	0–10	.9	0–0	.0
9. Mental ability	0–10	.9	0–4	.7	0–10	.8	0–5	.5
10. Personality of student	0–12	.8	0–2	.4	0–10	1.2	0–2	.1
11. Improvement shown	0–50	5.4	0–15	6.2	0–50	5.8	0–100	15.8
12. Consistency in preparation	0–25	5.3	0–3	.3	0–25	4.2	0–10	6.1
13. Whether or not a repeater	0–10	.4	0–10	3.2	0–5	.2	0–0	.0
14. Other factors	0–15	.2	0–0	.0	0–25	.7	0–10	.8

pupils rate themselves on these traits periodically. Conferences then follow with the pupils whose ratings of themselves differ significantly from those of the teachers.

The most important marks are those which are received at the end of the course and which become a part of the pupils' permanent record. How should these be determined? What should be the relative weight of the final examination, class tests, daily recitations, written work done outside the class, and the like? Unfortunately, there is no objective manner of determining the optimum weights. It must be a matter of judgment. The main thing is to obtain a reasonable degree of uniformity among the faculty. Such wide diversity of practice as is represented in Table 37 must be avoided.

Table 37

THE RELATIVE PROPORTION OF A COLLEGE STUDENT'S COURSE MARK IN MATHEMATICS AND ENGLISH DETERMINED BY DAILY RECITATIONS, QUIZZES AND SPECIAL REPORTS, AND FINAL EXAMINATIONS AS REPORTED BY TEACHERS IN THE NORTH-CENTRAL AND SOUTHERN ASSOCIATIONS OF COLLEGES AND SECONDARY SCHOOLS (after GILLIS)

Factor	Mathematics				English			
	N-Central		Southern		N-Central		Southern	
	Range	*Mean*	*Range*	*Mean*	*Range*	*Mean*	*Range*	*Mean*
1. Daily recitations	0–66	30.2	0–67	32.6	0–80	32.5	0–65	29.2
2. Quizzes and special papers	0–100	36.5	0–75	31.7	0–90	37.4	10–75	34.2
3. Final examinations	0–100	31.7	0–50	34.0	0–70	30.3	10–100	35.0
4. Other factors	0–50	2.0	0–25	1.5	0–35	.5	2–33	1.3

Before deciding upon a policy which is to govern the teaching staff, due consideration should be given to such points as the following: It is better, as a rule, to rely upon objective evidence rather than upon opinion, and upon written records rather than upon the teacher's memory. For this reason, except in the lower grades, many short tests afford a surer basis than the teacher's subjective impressions of daily recitations, oral quizzes, and the like. Work done out of class should receive less weight than that done in class, because of the difficulty of knowing the amount of assistance that may have been received. Exempting pupils from final examinations tends to create an unfavorable attitude toward measurement, removes a potent stimulus for a general review, and denies the superior pupil opportunity for developing skills he will find essential when he goes to college. On the other hand, allowing the final examination too much weight tempts pupils to trust last-minute cramming rather than regular, consistent study from day to day. Although average practice is never a safe guide of what ought to be, it is worth noting that Billett found that the median course mark in his 258 superior secondary schools was made up as follows: "Daily progress records, 40 per cent; term marks, tests given during semester, and the final examination, each 20 per cent." [26]

Definition of marks used. What do the various marks employed by the school really mean? Most teachers are embarrassed when asked to explain just what they mean by a mark of *A*, for example. The distinctions between marks are usually couched in language that is vague and confusing. This is probably one of the principal reasons for the diversity of practice in assigning marks.

Marks are usually thought of as being either absolute or relative. The percentage system is absolute. It has the appearance of being extremely simple, but in reality is subject to serious misinterpretation. A mark of 100 per cent does not mean that the pupil is perfect in the course, and 0 per cent does not mean the complete absence of knowledge. At best, the marks can mean only that the pupil was able to answer to the satisfaction of the teacher a certain percentage of the questions asked. Furthermore, such a system attempts a degree of refinement in educational measurement that is impossible of attainment today with the instruments available. In recent years there has been a definite trend away from the percentage system.

At the present time most marking systems are on a relative basis. Many schools use only two marks: *Pass* and *Failure,* or *Satisfactory* and *Unsatisfactory.* The former system is curriculum centered, while the latter is pupil centered. The pupil is marked either *Pass* or *Failure* according to the degree to which he has mastered the prescribed content of the course; or else he is marked *Satisfactory* or *Unsatisfactory* according to whether or not his educational progress has been consistent with his capacity. The latter is more in keeping with the spirit of the modern school. In fact, there is much to commend it, especially for reporting to parents. The confidential school records, however, should also take into account the individual's progress in relation to others, as well as in relation to his own capacity and past record, if he is to be guided intelligently in his subsequent educational and vocational choices. This will require a supplementary set of marks which recognize finer distinctions than merely *Satisfactory* or *Unsatisfactory.*

By far the most common system of reporting marks takes the form of letters, usually five in number. Billett found that four fifths of the secondary schools he studied issued marks "in the form of letters or equivalent symbols such as Arabic or Roman numerals."[27] At times an elaborate attempt is made to define each mark in terms of certain criteria. For example, a mark of *A* might be defined as follows:

---

[26] Roy O. Billett, *op. cit.,* page 438.

1. *Preparation:* Methodical, constant, exceeding expectations.
2. *Application:* Attention constant and concentrated; shows initiative.
3. *Knowledge of the subject:* Full and comprehensive, exceeding expectations.
4. *Use of English:* Extensive vocabulary, excellent diction, and so forth.
5. *Progress:* So rapid as to make the pupil an outstanding member of the group.

Such an attempt is not desirable, for several reasons. It is not only too complex and time-consuming for practical use, but it is also theoretically unsound. It includes factors other than achievement, some absolute and some relative, and it is highly subjective.

Most marking systems that make any pretense of being scientific are based upon the normal curve of probability. That there are no absolutely fixed percentages demanded by the normal curve is indicated by the fact that at least ten different distributions, each providing a 5-point system, have been defended by educators of distinction. These are given in Table 38. About as good a case can be made for one as for another, but one suggestion by Cajori in 1914 has the advantage of being most widely used. It also has the merit that the average values of the letter grades differ from each other by equal amounts. This is clearly shown in Figure 44. Whichever one is adopted, however, should be adhered to consistently.

Practice also varies on one other point. Some schools, perhaps the majority, think of each letter as representing a certain relative position *in the class of which the pupil is a member.* That is, when the pupil receives an *A,* he is in the highest 7 per cent of his class; when he receives a *B,* he is in the next 24 per cent of his class, and so on. On the other hand, it is sometimes suggested that the mark received should indicate the pupil's status *in relation to pupils in general of the same age or grade.* For example, if his achievement is 1.0 year or more above that of normal pupils, he is considered *A;* if he is between .5 and 1.0 year above normal, he is considered *B;* if he is within .5 of a year of normal, he is merely average, or *C;* similarly, if he is .5 to 1.0 year below normal, he is considered *D;* and if he is 1.0 year or more below normal he is considered *E.* There are two practical limitations to this system. In the first place, national

---

[27] *Ibid.,* page 426.

TABLE 38

PERCENTAGE DISTRIBUTIONS OF LETTER GRADES SUGGESTED BY VARIOUS WRITERS

E	D	C	B	A	WRITER	YEAR
10	20	40	20	10	Cattell	1905 [a]
3	22	50	22	3	Meyer	1905 [b]
2	23	50	23	2	Dearborn	1910 [c]
10	15	50	15	10	Smith	1911 [d]
4	24	44	24	4	Ruediger	1914 [e]
7	24	38	24	7	Cajori	1914 [f]
5	20	50	20	5	Brooks	1915 [g]
3½	24	45	24	3½	Rugg	1917 [h]
6	25	38	25	6	Ruch	1929 [i]
6	22	44	22	6	Eells	1930 [j]

[a] *Popular Science Monthly,* 14: 283.
[b] *Science, N. S.,* 28: 243.
[c] *University of Wisconsin Bulletin,* No. 368.
[d] *Journal of Educational Psychology,* 2: 385.
[e] *Science, N. S.,* 40: 643.
[f] *School Science and Mathematics,* 14: 283.
[g] *School and Society,* 1: 134.
[h] *Statistical Methods Applied to Education,* pages 216–219.
[i] *The Objective or New-Type Examination,* pages 378 ff.
[j] *Journal of Educational Psychology,* 21: 128.

norms are rarely available for the final composite score which determines the pupil's mark. Even if norms were available, the marks in many schools would not be well distributed; superior schools would have an undue concentration of *A's* and *B's,* and inferior schools a similar concentration of *D*'s and *E*'s.

There is a third possibility, which, in the judgment of the author, is best of all. According to this plan, each mark represents a given area *in the school or school system of which the pupil is a member.* Crawford states that there is "essential agreement" as to the following principles: [28]

> . . . (1) that the performance of a representative group sets the most logical standard of scholastic achievement; (2) that any individual's "mark" should properly represent his relative achievement within such a group; and (3) that equal relative achievement, as thus judged, should receive equal credit, whether in the same or different subjects of study.

According to this system, a pupil who receives an *A* is thought of as being in the highest 7 per cent of a typical group in his particular school or school system. In like manner, *B* indicates the next 24 per cent, *C* the middle 38 per cent, *D* the next lower 24 per cent, and *E* the lowest 7 per cent of the typical class.

---

[28] Albert Beecher Crawford, "Rubber Micrometers," *School and Society,* 32: 233-234, August 16, 1930.

It must be emphasized that *E* merely indicates that the pupil's performance places him in the lowest 7 per cent of a typical group in his school, *not necessarily failure.* Whether or not he is to be considered failing depends upon the promotional policy of the school and upon the teacher's educational philosophy. It cannot be too strongly emphasized that a rational curve system does not predestine any particular percentage to failure. Davis [29] has pointed out that over half the freshman class would be eliminated by the end of the senior year if 6 per cent were consistently failed each quarter for four years. In the elementary school and in non-vocational subjects in high school the primary consideration is the welfare of the pupil himself. In the long run, will he benefit by repeating the subject or grade, or by advancing with the group? In vocational subjects the standards required for admission to the vocation must be taken into account. In professions such as medicine and teaching the public welfare is the primary consideration in determining whether or not the achievement is satisfactory.

Summary of marking policy. The marking policy should be a group policy in the determination of which all members of the staff have a voice. The use of marks for motivation, for classification and promotion, and for informing parents of pupils' progress should receive less emphasis, and the guidance function should receive more. Since school marks are supposed to represent scholarship, only those factors should be considered which afford definite evidence of pupil achievement, and all extraneous items should be rigorously eliminated. Separate marks should be entered for other objectives regarded as important by the school. Each mark used should be carefully defined. Each letter mark should be thought of as describing a certain area in a group that is typical of the school or school system.

C. Essentials of a Sound Marking Technique

Having formulated a rational philosophy of marking, the school should next devise a suitable technique for putting the plan into actual operation. The marking technique should be adequate and at the same time not too complicated or unwieldy.

Determining the pupils' relative position or rank. The first step is a sound technique for determining the relative position of the various pupils in the class. This step is basic whether the mark is based on a single test or is a final mark at the end of the course. To determine these ranks, valid and reliable measuring instruments must be used. Perfectly valid and reliable tests would, of course, place each pupil in his exact order of merit, but such an ideal can be only approximated. While all measurement is subject to error, the errors may be kept within reasonable limits if the teachers understand the functions and limitations of each type of measuring instrument employed and have developed the skill necessary for its construction and use. No marking plan can be successful without this skill. The marks can never be more valid than are the scores or ranks upon which they are based.

---

[29] J. De Witt Davis, "The Effect of the 6–22–44–22–6 Normal Curve System on Failures and Grade Values," *Journal of Educational Psychology,* 22: 636–640, November, 1931.

Transmuting scores into marks. In an absolute system there is no distinction between a score and a mark. The pupil's score is usually represented as a certain percentage of perfection, and that is all there is to it. In a relative system, however, the score is merely a numerical description of a pupil's performance. In order to be meaningful, it must still be transmuted into the appropriate unit recognized by the marking system of the school. This derived score is called a *mark.*

Determining "satisfactory" and "unsatisfactory." In certain school systems a pupil is marked *Satisfactory* if his achievement is regarded as consistent with his learning capacity, and *Unsatisfactory* if it is not. There is much to be said for this system, especially as used for reporting to parents. It will be recalled that in the Biblical parable the two-talent man and the five-talent man received exactly the same recognition: "Well done, thou good and faithful servant." But it is desirable to determine the mark as objectively as possible. McCall [30] makes the suggestion that a pupil should be marked *Unsatisfactory* only when his achievement falls one full year behind his intelligence. Even then he should not be so marked unless he is known to be free from any serious physical handicap and is properly classified in school. Adams [31] proposes a somewhat more complicated scheme for the same purpose. He would express the pupil's "working ability" in deciles according to a composite rating made up of his intelligence test score, the teachers' estimate of his ability, his health rating by the nurse or doctor, and his home rating by a visiting teacher, school nurse, or attendance officer. He would express the pupil's achievement in deciles according to a composite rating made up of his scores in a series of objective tests and his rank in accomplishment assigned by the teacher "regardless of

---

[30] William A. McCall, *Measurement,* pages 446–447. New York: The Macmillan Company, 1939.

[31] A. Elwood Adams, "Marking Pupils on Their Working Ability," *Nation's Schools,* **17**: 35–36, April, 1936.

. . . intelligence, home conditions, health, and teacher opinion of ability." The pupil is then marked *Unsatisfactory* if his decile score in achievement falls two or more points below that of his "working ability." It can be seen that such a scheme is somewhat subjective, highly arbitrary, time-consuming, and makes no attempt to distinguish different degrees of satisfactoriness and of unsatisfactoriness that may be desirable for guidance. This system should be attempted only where there is a highly trained staff and adequate clerical assistance.

Transmuting point scores into letter marks. The simplest scheme of transmuting point scores into letter marks is merely to count off the highest 7 (or other similar) per cent for *A*, the next highest 24 per cent for *B*, and so on. But this system is not sufficiently flexible. It takes no account of differences among classes in central tendency or in variability. This is the primary weakness of most so-called curve systems. It is probably fairly safe to assume that successive grades in the elementary school are very much alike, but this is far from true of different classes and departments in high school and college.[32] And the extent of such differences should be *objectively determined* and not merely estimated. The accurate determination of the status of the class is fundamental to any equitable distribution of marks.

An attempt is often made to provide the needed flexibility by suggesting ranges within which each letter grade may fall. The following is a sample:

A is the highest 5 to 10 per cent.
B is the next highest 20 to 30 per cent.
C is the middle 35 to 45 per cent.
D is the next lowest 20 to 30 per cent.
E is the lowest 5 to 10 per cent.

Such a policy does nothing to provide a basis for determining the proper proportions within these ranges that are most defensible for a particular class. Some schools attempt to meet this need by suggesting to the teacher a reasonable distribution of marks to be expected, based upon the ability of the class as indicated by tests of general intelligence or of reading ability, or by total academic standing to date. Spence[33] and McCall[34] propose a division of responsibility between the classroom teacher and some central administrative office. The teacher merely indicates the relative *ranks* of the students, and the office converts them into *marks* according to the distribution of intelligence in the class. This plan, however, is open to the theoretical objection that the correlation between achievement and other factors is far from perfect, and to the practical objection that the scheme involves more labor than most schools will be willing to devote to it.

---

[32] *Cf.* William F. Book, *The Intelligence of High School Seniors,* page 146. New York: The Macmillan Company, 1922; Karl C. Pratt and Virgil Wise, "Do Marking Systems Based upon the Normal Probability Curve Insure an Equitable Distribution of Marks in Elective Curricula?" *Journal of Experimental Education,* 5: 261–270, March, 1937.

[33] Ralph B. Spence, *The Improvement of College Marking Systems,* pages 64–71, New York: Bureau of Publications, Teachers College, Columbia University, 1927; also, *School and Society,* 28: 224–227, August 25, 1928.

[34] William A. McCall, *op. cit.,* page 468.

Table 39

RELATION BETWEEN CUMULATIVE POINT STANDING AND CLASS MARKS FOR TWENTY-ONE COLLEGE CLASSES
(AFTER LEKER)

Class	Number Students	Cumulative Standing	Class Marks	Correlations
Spanish	31	1.49	1.26	.65
Bible Poetry	11	1.48	1.64	.69
Educational Psychology	30	1.47	1.47	.67
Political Geography	10	1.46	1.40	.38
Mental Hygiene	11	1.43	1.54	.66
Latin American History	21	1.43	1.20	.48
Trigonometry	12	1.42	1.33	.67
Evaluation	14	1.41	1.43	.53
Grasses	4	1.41	2.00	.55
History of Philosophy	7	1.40	1.71	.52
History of England	16	1.39	1.37	.60
Chemistry 1	25	1.38	1.64	.59
Botany 1	28	1.35	1.04	.52
Social Trends	14	1.32	1.21	.64
Physiology	9	1.25	1.55	.35
Personality	18	1.24	1.39	.65
Physics 1	20	1.24	1.20	.11
Handicrafts	19	1.23	1.53	.49
Hygiene	33	1.19	1.18	.58
Spanish Literature	29	1.19	1.28	.60
Lettering	15	1.18	1.33	.50
Median	18	1.39	1.39	.58

Leker [35] has reported an instructive experiment which attempted to regulate the distribution of college marks. Shortly before the final marks are to be reported each instructor receives the cumulative point standing to date of the students in his class. It is suggested that in classes for juniors and seniors the average mark should not vary up or down more than .25 of a point from this point standing. Table 39 shows the results for the first summer term of 1944.

---

[35] Charles A. Leker, "Previous Class Cumulative Index as a Guide to Grading," *Journal of Educational Research,* 39: 56–61, September, 1945.

Several points are worthy of note. Variations in ability appear to bear little relationship to the school subject or the class size. The correlations between cumulative standings and class marks which are shown in the column at the right were generally higher than those usually reported between intelligence test scores and class marks. It is also clear that correlations may be fairly high even when the average class mark is considerably higher or lower than the average point standing, or the correlations may be fairly low when the averages agree clearly. Apparently the scheme has by no means eliminated all inconsistencies. For example, the marks in the two Spanish classes fail to reflect the differences in ability.

The technique for transmuting point scores into letter marks should be sufficiently flexible to accommodate itself to differences among groups both in variability and in central tendency. Unfortunately all such techniques appear to be rather complicated. In reality, however, they are much simpler than they appear to be upon first acquaintance. All such schemes are based upon the central tendency and the variability of the group. The mean (M) is used as a measure of central tendency, and either the standard deviation (σ) or the mean deviation (MD) is used as a measure of variability. In a five-letter system based upon M and σ, the areas representing the various letters are described as follows:

A is 1.5σ or more above the mean.
B is between $+.5\sigma$ and $+1.5\sigma$.
C is between $-.5\sigma$ and $+.5\sigma$.
D is between $-.5\sigma$ and -1.5σ.
E is 1.5σ or more below the mean.

The various areas in a normal distribution indicated by the various letter marks are given in Fig 44.

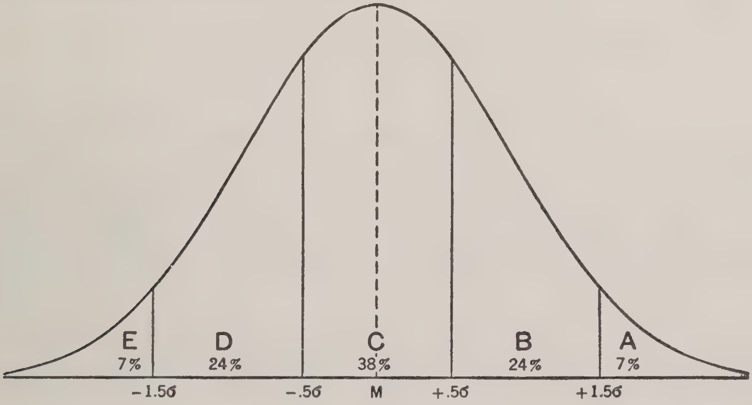


Figure 44. Areas in a Normal Distribution, Indicated by a Five-Letter Marking System Based on M and σ.

In a five-letter system based upon M and MD, the areas representing the various letters are described as follows:

A is $2MD$ or more above the mean.
B is between $+\frac{2}{3}MD$ and $+2MD$.
C is between $-\frac{2}{3}MD$ and $+\frac{2}{3}MD$.
D is between $-\frac{2}{3}MD$ and $-2MD$.
E is $2MD$ or more below the mean.

It will be observed that the areas representing each letter are uniform in width, either 1σ or $1\frac{1}{3}MD$.

Table 40 illustrates the two procedures for a class of 30 pupils. Although the scheme based on M and σ involves somewhat more work, it is used more extensively than the other, but there is no good reason why it should be. The limits for the letter marks differ slightly in the two procedures, but the letter distributions usually agree very closely. In this situation they are identical. Both procedures give 2 *A*'s, 6 *B*'s, 13 *C*'s, 7 *D*'s, and 2 *E*'s.

Except in small classes, it is usually better to make a frequency table and then to compute M and σ by the short method. This procedure is illustrated in Table 41. The number of individuals entitled to the various letter marks can be determined only approximately from the table, but can be found exactly by referring to the actual scores in the teacher's class record book.

Jenkins[36] has proposed a short-cut method for estimating the standard deviation which appears to be sufficiently accurate for the purpose of distributing marks. The three simple steps required for making the estimate will be illustrated by the data in Table 44.

1. Find the mean of the highest 10 per cent of the scores. Mean of 160, 154, and 145 is 153.
2. Find the mean of the lowest 10 per cent of the scores. Mean of 64, 56, and 49 is 56.3.
3. Divide the difference between these means by 3.4.

$$\frac{153 - 56.3}{3.4} = 28.4$$

Note that the estimate differs from the computed value by only .2, a difference so small as not to affect the letter value of a single score.

In the first illustration there is a close agreement with the typical 7–24–38–24–7 distribution. Even here, however, there is a slight departure: the *B*'s fall below expectation by 4 per cent, and the *C*'s exceed expectation by 5 per cent. That the system permits much greater flexibility than this is illustrated by Table 41. In this class there are fewer *C*'s than *B*'s and four times as many *E*'s as *A*'s. This is because of the form of the distribution, which is negatively

---

[36] William Leroy Jenkins, "Short-Cut Method of Estimating Standard Deviations," *American Psychologist,* 1: 247, July, 1946.

Table 40

TWO TECHNIQUES FOR TRANSMUTING POINT SCORES INTO LETTER MARKS

Point Scores	Deviations from Mean	Squares of Deviations	Transmutation into Marks
160	+56	3,136	M is the sum of point scores divided by N: $3{,}117 \div 30 = 103.9$
154	+50	2,500	
145	+41	1,681	MD is sum of deviations from M divided by N (Note that M is taken at nearest integer, 104, and that the deviations are added without regard to sign): $699 \div 30 = 23.3$
142	+38	1,444	
140	+36	1,296	
133	+29	841	
130	+26	676	
128	+24	576	
117	+13	169	σ is square root of the mean of the deviations squared:
114	+10	100	
112	+8	64	
112	+8	64	$23{,}873 \div 30 = 795.8$; $\sqrt{795.8} = 28.2$
110	+6	36	
108	+4	16	If A's start $2MD$ above M, and other letters are separated by $1\frac{1}{3}$ MD, the limits of the marks[a] are:
100	−4	16	
98	−6	36	
97	−7	49	
96	−8	64	150 and up = A
94	−10	100	119 to 149 = B
93	−11	121	88 to 118 = C
90	−14	196	57 to 87 = D
86	−18	324	56 and below = E
85	−19	361	
82	−22	484	If A's start 1.5σ above M, and other letters are separated by 1.0σ, the limits of the marks[a] are:
77	−27	729	
75	−29	841	
72	−32	1,024	146 and up = A
64	−40	1,600	118 to 145 = B
56	−48	2,304	89 to 117 = C
49	−55	3,025	61 to 88 = D
30)3,117	30)699	30)23,873	60 and below = E
103.9	MD = 23.3	795.8	
		$\sqrt{795.8} = 28.2$	[a] The decimals are dropped.

skewed rather than normal. It is apparent that this system, instead of imposing a fixed percentage for each letter, readily adjusts itself to the shape of the distribution. This flexibility is a decided advantage.

Allowing for variations in central tendency. If the school has adopted the policy of defining a pupil's mark in terms of the position he occupies *in his own class*, there is no further problem. But if the school has adopted the more defensible policy of defining a pupil's mark in terms of his position *in a typical class in the school,*

TABLE 41

TRANSMUTING POINT SCORES INTO LETTER GRADES FROM A FREQUENCY TABLE

COMPUTATION OF M AND σ [a]

Score	f	d	fd	fd²
220–	3	+5	15	75
210–	5	+4	20	80
200–	8	+3	24	72
190–	8	+2	16	32
180–	6	+1	6	6
170–	4			
160–	5	—1	— 5	5
150–	3	—2	— 6	12
140–	4	—3	—12	36
130–	2	—4	— 8	32
120–	1	—5	— 5	25
110–	0	—6	0	0
100–	0	—7	0	0
90–	1	—8	— 8	64

$N = 50$ | fd: +81, —44; $50\overline{)\,+37}$ | fd²: 439

$M^1 = 175$
$ci = +7.4$
$M = 182.4$

$c = .74$
$c^2 = .5476$

$$\sigma = 10\sqrt{\frac{439}{50} - (.74)^2}$$

$$= 10\sqrt{8.78 - .5476}$$

$$= 10\sqrt{8.2324}$$

$$= 10 \times 2.87 = 28.7$$

[a] For a detailed discussion of process, see Chapter VIII.

TRANSMUTATION INTO MARKS

Limits of letter marks: [b]

225 and up = A
196 to 224 = B
168 to 195 = C
139 to 167 = D
138 and below = E

Number and per cent of letters awarded:

Letter	Number	Per Cent
A	1	2
B	18	36
C	17	34
D	10	20
E	4	8

[b] Decimals are dropped.

the crucial problem remains of determining the status of his class. Is this particular class typical of the school? If not, what is the amount and direction of the departure?

If the class is in a subject required of all pupils, or if it is large, it will probably not be far from the average of the school. On the other hand, if the class is elective, or rather small, it may be anywhere from very superior to very inferior. The two smallest classes which receive the highest marks are of but little better than average ability as reflected in cumulative point standing to date. One of the smallest classes is below average ability and another is near the top. Likewise, two of the largest classes rank near the top in ability and two other larger classes rank near the bottom. In any case, it is better to determine the status of the class, rather than merely to assume it. But in the absence of objective evidence to the contrary, the safest assumption to make for any class is that it is probably about average for the school.

The status of a class is established by comparing its central tendency with that of representative classes on the basis of general intelligence, general academic standing, reading ability, or scores on standardized achievement tests. Table 42 represents the scores of three hypothetical classes on the same achievement test. If Y is a typical class, X is somewhat inferior and Z is somewhat superior, although all have exactly the same variability. The usual 7–24–38–24–7 distribution of marks is satisfactory for Y, but not for X and Z. Unless the differences in central tendency are taken into account, 47 per cent of the pupils in class X will receive higher marks than they deserve, and 47 per cent of those in class Z will receive lower marks than they deserve.[37] On the basis of equal achievement, for

Table 42

SCORES FOR THREE CLASSES WHOSE VARIABILITY IS THE SAME BUT WHOSE CENTRAL TENDENCIES DIFFER

Score	*X*	*Y*	*Z*	
130–			1	*A*
120–		1	2	
110–	1	2	4	
100–	2	4	9	
90–	4	9	15	*B*
80–	9	15	19	
70–	15	19	19	*C*
60–	19	19	15	
50–	19	15	9	*D*
40–	15	9	4	
30–	9	4	2	*E*
20–	4	2	1	
10–	2	1		
0–	1			
N	100	100	100	
M	60	70	80	
σ	20	20	20	

---

[37] So great are the differences among classes and schools that Davison found, from a state-wide study, that in algebra, for example, some pupils received *A*'s whose test scores were only 12, while other pupils in better schools were failed whose test scores were as high as 30. F. M. Davison, Master's Thesis, State University of Iowa, 1933.

example, instead of 7 A's each in classes X and Z, there should be only 3 in X and 16 in Z.

How can these differences in central tendency be taken into account? The technique for doing this is simple. It will be recalled that A has been defined as an area located 1.5σ or more above the mean of a typical class. The problem is to determine the mean of a typical class and then to locate the letter marks with reference to this mean rather than to the mean of the class itself, which is non-typical. In Table 42 the mean of the typical class, Y, is 70. Therefore, all pupils in classes X and Z who are 1.5σ or more above 70, the mean of a typical class, receive A. In like manner, those between $.5\sigma$ and 1.5σ receive B, and so on. Since σ is 20 in all classes, the distribution of letter marks is as follows:

Score Range	Letter Mark	Number in Each Class		
		X	*Y*	*Z*
100 and above	A	3	7	16
80 to 99	B	13	24	34
60 to 79	C	34	38	34
40 to 59	D	34	24	13
39 and below	E	16	7	3

If standard tests have been used throughout the school, the mean of these scores is that of a typical class in the school. If standard test scores are not available, the mean mark, or standing, in the school may be used. The status of any class is then determined by the amount its mean exceeds or falls behind the mean of the typical class. For example, suppose that the mean IQ of a high school is 105, and that of a particular class in this school is 110, with a σ of 10. This comparison shows that the average ability of this class is $.5\sigma$ above that of the typical class.

Now suppose that at the end of the semester the mean composite score of this class on all measures of achievement used is 200, with a σ of 30. Since this class exceeds the typical class in intelligence by $.5\sigma$, the best estimate than can be made is that its mean achievement score likewise exceeds that of the average class by $.5\sigma$, or 15 points. The A's should then start at 1.5σ above 185, the mean of the typical class in the school, rather than 1.5σ above 200, the mean of this superior class: that is, $(200–15) + (1.5 \times 30) = 230$, minimum score for A. Marks B, C, and D would begin at 200, 170, and 140, respectively. Likewise, if the mean intelligence score of the class had been 100, or $.5\sigma$ below the school average, the mean achievement score of the typical class would have been estimated at 215. In that case, *A* would have begun at 260, *B* at 230, *C* at 200, and *D* at 170.

Let us refer again to Table 39 on page 414. Observe that one Spanish class, with a mean cumulative point standing of 1.49, is .10 above the mean [38] of the typical class which is 1.39. On the other hand, the class in Spanish Literature and the Hygiene class are each .20 below the mean of the typical class. If the σ is .2, one Spanish class ranks .5σ *above* the mean of the typical class in the college and the other Spanish class is 1σ *below* the mean of the typical class. Manifestly this difference has been completely ignored in the distribution of marks.

It is also possible to combine the scores on several sections or classes that have had the same test, whether standardized or not. The mean of this combination is likely to be very close to that of an average class in this subject. The deviation of any particular class from this average indicates whether it is an inferior, a superior, or just an average class. If this deviation is divided by the σ of the class, a standard score is obtained, which affords a basis for estimating the probable value of the mean of a typical class on any future test or combination of tests. For example, suppose that the teacher makes out an entirely new final examination and gives it to a class which the above procedure indicates is about .3σ below the average class in the subject. If the class mean on this test is 90 and the σ is 12, the mean of a typical class on this test would probably be about 3.6 points higher, or 93.6. The letter marks can then be determined from this estimated mean. The result would have been fewer *A*'s and *B*'s and more *D*'s and *E*'s than would be found in a typical class, if the evidence had not indicated that the class was somewhat below the average for the subject.

The point to stress is not the use of any particular technique in making the needed adjustment for classes of unequal ability, but rather that such differences do exist and should be taken into account in distributing marks. No so-called "curve" system of marking can make any pretense of being scientific which ignores the differences *among* classes any more than the differences *within* classes. Experienced teachers often learn to make, on a subjective basis, adjustments for differences in ability which are reasonably satisfactory.

Summary of marking technique. The first requirement is a sound technique for determining the correct rank order of the pupils in the class, usually in terms of point scores. This ranking presupposes the use of measuring instruments that are as valid and as reliable as possible. The second requirement is a satisfactory system for transmuting these ranks or scores into the marks recognized by the school. If marks of *Satisfactory* and *Unsatisfactory* are used, both achievement and capacity should be determined as objectively as possible, and definite standards should be set up as to how much a pupil's achievement must fall behind his capacity before his progress is regarded as *Unsatisfactory.* A practical scheme for transmuting point scores into letter grades is in terms of the mean and standard deviation of the distribution. A plan has been suggested for making allowance for classes which depart from the normal or typical class in the school. Reasonable provision for nontypical classes is a crucial matter in any curve system.

---

[38] It is here assumed that the mean is the same as the median.

SELECTED REFERENCES FOR FURTHER READING

Billett, Roy O., *Provisions for Individual Differences, Marking and Promotion.* National Survey of Secondary Education, Monograph No. 13. Washington, D. C.: United States Office of Education, 1932. Part IV.

Elsbree, Willard S., *Pupil Progress in the Elementary School.* New York: Bureau of Publications, Teachers College, Columbia University, 1943. Chapter VII.

Hawkes, Herbert E., Lindquist, E. F., and Mann, C. R., *The Construction and Use of Achievement Examinations.* Boston: Houghton Mifflin Company, 1936. Pages 118–125.

Jacobson, Paul B., and Reavis, William C., *Duties of School Principals.* New York: Prentice-Hall, Inc., 1941. Chapter XIV.

Lee, J. Murray, *A Guide to Measurement in Secondary Schools.* New York: D. Appleton-Century Company, 1936. Chapter VII.

McCall, William A., *Measurement.* New York: The Macmillan Company, 1939. Book Six.

Odell, C. W., *Educational Measurement in High School.* New York: D. Appleton-Century Company, 1930. Chapter XIX.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman & Company, 1929. Chapter XIV.

Spence, Ralph B., *The Improvement of College Marking Systems.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 89 pages.

Symonds, Percival M., *Measurement in Secondary Education.* New York: The Macmillan Company, 1927. Chapter XXIV.

Tiegs, Ernest W., *Tests and Measurements in the Improvement of Learning.* Boston: Houghton Mifflin Company, 1939. Chapter XVII.

Williams, L. A., *Secondary Schools for American Youth,* New York: American Book Company, 1944. Chapter X.

CHAPTER XV

Classification and Promotion

A. The Nature and Educational Significance of Human Variability

The problem of human variability. The existence of variability is one of the best established facts about human beings. Obvious differences in height, weight, strength, and good looks could hardly escape the notice of the most casual observer. The greatest seers and wise men of all ages have recognized also the less obvious but more important differences in ability, interests, and needs. One of the familiar parables of Jesus, for example, is that of the talents.[1]

It would be difficult today to find a fuller recognition of the educational significance of individual differences than appears in the writings of those two apostles of human liberty, Jean Jacques Rousseau and Thomas Jefferson. Rousseau asserted that "it would be a great mistake to bestow it (instruction) on all children indiscriminately and without regard to their individual differences."[2] Jefferson wrote of a proposed educational measure: "The general objects of this law are to provide an education adapted to the years, capacity and the condition of every one, and directed to his freedom and happiness."[3] It is apparent, therefore, that when the author of the Declaration of Independence penned the famous line, "All men are created equal," he had in mind *equality before the law,* and that he recognized fully the duty of the state through education to provide *equality of opportunity.* A prominent American educator[4] argues that "our concepts of freedom and equality are outmoded" and cannot both be realized, since they "are in fact mortal enemies."

It is surprising, therefore, to find that the problem of individual differences was not seriously treated in psychology before the time of Galton in the latter half of the nineteenth century, a neglect which has been characterized as perhaps the "most extraordinary blind-spot in previous psychology."[5]

---

[1] "And unto one he gave five talents, to another two, and to another one; to every man according to his several ability." Matthew 25:15.

[2] Jean Jacques Rousseau, *The New Heloise,* Part V, Letter 3.

[3] Thomas Jefferson, *Notes on Virginia,* pages 250–252.

[4] I. Newton Edwards, "We Need New Purposes in Education," *Phi Delta Kappan,* 28: 16, September, 1946.

[5] Gardner Murphy, *An Historical Introduction to Modern Psychology,* page 123. New York: Harcourt, Brace & Company, Inc., 1930.

Otto[6] estimates that during the last twenty years more time and effort in educational research have been devoted to the study of individual differences than to any other single topic; he is greatly impressed with the extensive literature available. Yet in 1925 a competent school psychologist stated that the "schools heretofore have to a large extent ignored these differences."[7] Five years later a national survey revealed that "provisions for individual differences, in general, are innovations in the secondary schools."[8] About a decade ago a survey of 300 courses of study showed that only about one in ten "contain any suggestions for adapting instruction to individuals."[9] Recently an educational psychologist[10] characterized as largely "lip service" the attention educators give to individual differences. Davis says:

> Despite its philosophy of individualization, the school, in practice, fosters a program of regimentation and standardization.

In the meantime, better enforcement of the compulsory education laws and the rapid increase of secondary-school enrollments have served but to intensify the problem, the nature of which was more accurately revealed by scientific measurement.[11]

Group differences. Scientific research, on the whole, has shown that differences between groups are not so great as they are commonly assumed to be. There is little basis for the widespread illusion that the group of which one happens to be a member is superior, while all others are inferior. The intellectual differences between the sexes, for example, are slight. Furthermore, all levels of mental ability are found in all economic, occupational, and social groups, although not in the same proportions. Even the differences between races have been grossly exaggerated, and such differences as appear reflect cultural rather than innate intellectual variations. It is manifestly impossible to make adequate provision for individual differences by classifying pupils for instructional purposes according to the social, economic, occupational, racial, or other similar group from which they come. Fortunately, perhaps, for democracy the problem is not so simple as that.

---

[6] Henry J. Otto, *Elementary School Organization and Administration* (Second Edition), page 160. New York: D. Appleton-Century Company, 1944.

[7] A. A. Sutherland, "Factors Causing Maladjustment of Schools to Individuals," *Twenty-Fourth Yearbook of the National Society for the Study of Education, Part II,* pages 29–30. Bloomington, Illinois: Public School Publishing Company, 1925.

[8] Roy O. Billett, *Provisions for Individual Differences, Marking and Promotion,* National Survey of Secondary Education, Monograph No. 13, page 8. Washington, D. C.: United States Office of Education, 1932.

[9] Henry Harap, "Differentiation of Curriculum Practices and Instruction in Elementary Schools," *Thirty-Fifth Yearbook of the National Society for the Study of Education, Part I,* page 162. Bloomington, Illinois: Public School Publishing Company, 1936.

[10] Robert A. Davis, "Experimenting in Education," *Educational Administration and Supervision,* 30: 1–16. January, 1944.

[11] For an excellent recent summary, see A. R. Gilliland and E. L. Clark, *Psychology of Individual Differences,* 535 pages. New York: Prentice-Hall, Inc., 1939.

Almost without exception the average differences between groups are less significant than the differences within any single group. An important example of this is the enormous overlapping among school grades. Although the average difference in intelligence and in achievement between successive school grades rarely exceeds one year, the difference within any grade is likely to be at least four or five years. As a matter of fact, Baker [12] points out that the achievement of the more capable halves, or the less capable halves, of two adjacent grades is usually much more alike than that of the two halves of the same grade. On a test of general academic knowledge, the Pennsylvania study showed that about 10 per cent of the high-school seniors exceeded the median of the college seniors, while nearly 10 per cent of the college seniors fell below the median of the high-school seniors.[13]

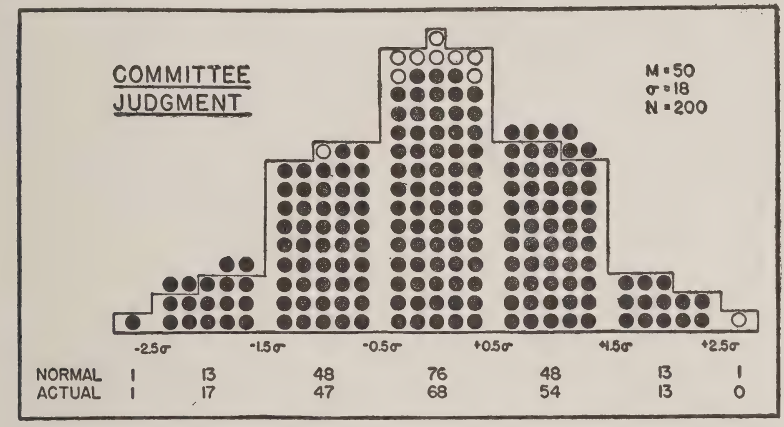


Figure 45. General Quality of 200 Secondary Schools as Judged by Field Committees. Each ● represents an actual school. Each ○ represents a school in theoretical distribution. (From *Education of Secondary Schools, General Report,* page 110.)

It must not be thought, however, that one group is just like every other. As a matter of fact, certain types of groups differ from each other very much as the individuals within any one group differ from each other. For example, Figure 45 shows the distribution of the ratings of 200 high schools, which closely approximates the normal curve. It is quite likely that all the schools in a single state would be similarly distributed on practically every characteristic.

---

[12] *Thirty-Fifth Yearbook, op. cit.,* pages 137, 145.

[13] William S. Learned and Ben D. Wood, *The Student and His Knowledge,* page 21. New York: Carnegie Foundation for the Advancement of Teaching, 1938. For a recent illustration from the results of the Army General Classification Test, see: Walter V. Bingham, "Inequalities in Adult Capacity—from Military Data," *Science,* **104**: 147–152, August 16, 1946.

Figure 46, although somewhat skewed, shows that, on the basis of the mean achievement of their seniors, 49 colleges in Pennsylvania have the extremely wide range and the heavy concentration near the center that characterize normal curves. In other words, there are differences among institutions just as there are among individuals. It is this fact that makes the traditional classification of schools for accrediting purposes such a baffling problem, and that has been responsible for the trend toward evaluating each school in relation to its own objectives and program rather than in relation to other schools. It is now being recognized that it is just these differences that give individuality and distinction to institutions.

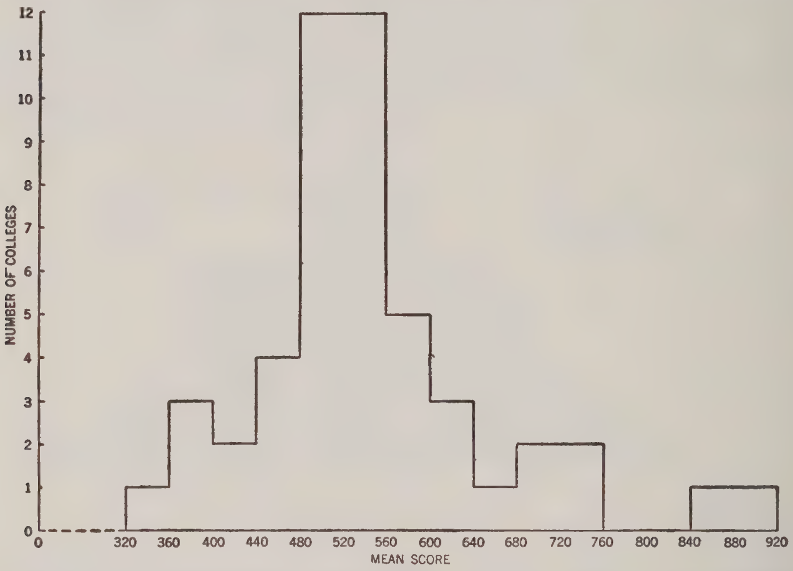


Figure 46. Distribution of Mean Scores of Seniors in Forty-Nine Colleges in Pennsylvania on a Test of General Academic Knowledge. (Data from *The Student and His Knowledge,* page 78.)

Individual differences. In contrast with the differences between groups, which have frequently been overestimated, the differences within the group have usually been underestimated. While a vague notion of individual differences has long been in existence, no adequate knowledge of the nature and extent of these differences was possible before the appearance of scientific measurement. Such profound thinkers as Plato, for example, believed that all persons fell into a few rather distinct groups. In fact, the idea that individuals are distributed according to the normal curve is a modern conception.

Figure 47, according to Terman and Merrill, "probably gives the clearest picture available of the intellectual differences which obtain among American-born white children of the ages in question:"[14] Figure 21 on page 267 shows a similar distribution for ninth-grade pupils. Three characteristics of the so-called "normal curve" should be noted: (1) the *wide range* from lowest to highest scores, (2) the *continuous distribution*—no breaks, and (3) the distinct *tendency to pile up near the center.* With only a few exceptions, curves representing all human traits have these same characteristics. Skewed curves differ from normal curves only in that the heavy concentration is not exactly at the center. The fact that the best pupil in any group on any test is likely to make a score from two to five or more times the score of the poorest has great educational implications, as also does the fact that approximately two thirds of the pupils lie within a standard deviation distance from the mean. After a survey of the experimental evidence, Hull says:[15]

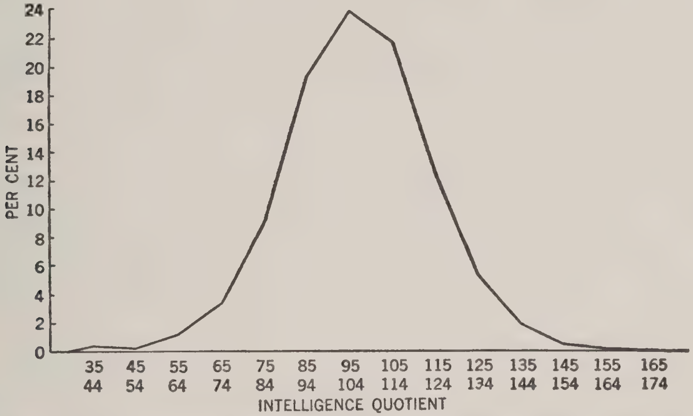


Figure 47. Distributions of Composite IQ's on Forms L and M of the New Revised Stanford-Binet Tests for a Standardized Group of 2,904 Individuals of CA's 2 to 18 years. (From Terman and Merrill *Measuring Intelligence,* page 37.)

> We shall probably not be in great error if we conclude that *among individuals ordinarily regarded as normal, in the average vocation the most gifted will be between three and four times as capable as the poorest.*

---

[14] Lewis M. Terman and Maud A. Merrill, *Measuring Intelligence,* page 37. Boston: Houghton Mifflin Company, 1937.

[15] Clark L. Hull, *Aptitude Testing,* page 36. Yonkers, New York: World Book Company, 1928.

Important as is the wide range of ability between the two extremes, the importance to education of the *continuous distribution* is equally great. On no trait do individuals naturally fall into a few distinct groups, such as "inferior," "average," and "superior," or "dull," "normal," and "bright." Such so-called "types" are purely arbitrary. It would be possible to make an equally good case for any other number of classes. "In a literal sense, everyone is exceptional."[16] There are similar differences in nonintellectual traits.

Trait variability. Not only are there differences among groups, and differences among the individuals of any one group, but there are also important differences among the traits making up any particular individual. Hull made a careful study of these differences and came to the conclusion that "the distribution of talent within an individual follows the normal law much as do the distributions of individual differences."[17] Not only did he observe a "distinct tendency to approach the characteristic shape of the normal probability curve," but he also found evidence that "the average individual's best vocational potentiality must be between two and one-half and three times as good as his worst."[18] The importance of these trait differences in an individual for educational and vocational guidance can hardly be overemphasized. It is also apparent that satisfactory ability grouping in one trait may be wholly unsatisfactory in other traits.

The educational problem is further complicated by the fact that, while intercorrelations of these factors are usually positive, the correlations are far from perfect. This means that when an attempt is made to secure a group homogeneous in one factor, it is still heterogeneous with respect to other factors. It is apparent, therefore, that it is impossible to make groups truly homogeneous for instructional purposes, even if it were desirable to do so. The best that can be done is to reduce the amount of heterogeneity. Opponents of ability grouping have made much of this point, apparently quite oblivious to the fact that *ipso facto* they are attacking a straw man; for, manifestly, one need shed no tears over the dangers of an educational situation which one's own data prove to be a physical impossibility.

The concept of the versatile individual who is equally gifted in a considerable number of directions is largely a fiction and as an educational ideal is capable of doing much harm. It has been ridiculed as follows by a capable scientist: [19]

---

[16] Edmund S. Conklin and Frank S. Freeman, *Introductory Psychology for Students of Education*, page 515. New York: Henry Holt & Company, 1939.

[17] Clark L. Hull, *op. cit.*, page 46.

[18] *Ibid.*, pages 46, 49.

> In antediluvian times, while the animal kingdom was being differentiated into swimmers, climbers, runners, and fliers, there was a school for the development of the animals.
>
> The theory of the school was that the best animals should be able to do one thing as well as another.
>
> If an animal had short legs and good wings, attention should be devoted to running, so as to even up the qualities as far as possible.
>
> So the duck was kept waddling instead of swimming. The pelican was kept wagging his short wings in the attempt to fly. The eagle was made to run, and allowed to fly only for recreation.
>
> All this in the name of education. Nature was not to be trusted, for individuals should be symmetrically developed and similar, for their own welfare as well as for the welfare of the community.
>
> The animals that would not submit to such training, but persisted in developing the best gifts they had, were dishonored and humiliated in many ways. They were stigmatized as narrow-minded and specialists, and special difficulties were placed in their way when they attempted to ignore the theory of education recognized in the school.
>
> No one was allowed to graduate from the school unless he could climb, swim, run, and fly at certain prescribed rates; so it happened that the time wasted by the duck in the attempt to run had so hindered him from swimming that his swimming muscles had atrophied, and so he was hardly able to swim at all; and in addition he had been scolded, punished, and ill-treated in many ways so as to make his life a burden. He left school humiliated, and the ornithorhynchus could beat him both running and swimming. Indeed, the latter was awarded a prize in two departments.
>
> The eagle could make no headway in climbing to the top of a tree, and although he showed he could get there just the same, the performance was counted a demerit, since it had not been done in the prescribed way.
>
> An abnormal eel with large pectoral fins proved he could run, swim, climb trees, and fly a little. He was made valedictorian.

Educational provisions for individual differences. Attention has already been called to the fact that few schools are making adequate provisions for the individual differences existing in their pupils. No point stood out more prominently in Billett's study [20] than this. Table 43 summarizes the situation for secondary schools in 1930. Billett reduces these provisions to seven categories: (1) homogeneous grouping, (2) special classes, (3) plans characterized by the unit assignment, (4) scientific study of problem cases, (5) variation in pupil load, (6) out-of-school projects and studies, and (7) advisory or guidance programs. Of these the first three "have been found to be core elements in a typically successful program to provide for individual differences." [21] But it will be noted from the last column that the most successful provision, in the opinion of those using it, is homogeneous grouping, which has a ratio of 26 per cent. In other words, hardly more than one principal in four or five using any of these plans has a considerable degree of confidence in them.

---

[19] Amos E. Dolbear, "Antediluvian Education," *Journal of Education,* 68: 424, 1908.

[20] Roy O. Billett, *op. cit.,* pages 8–11.

Table 44, based on returns from 48 large and 58 small cities, shows recent trends in the elementary school.[22] Increasing attempts to introduce more flexible educational programs are shown. It is also apparent that much yet remains to be done. But it is encouraging to note that somewhat more than a third of these schools are using reading-readiness tests and other tests for diagnostic and guidance purposes rather than merely as a basis for promotion.

B. Differentiated Unit Assignments

Many schemes of classification and instruction are used, all of which employ units or unit assignments in some form, but which differ from one another in other respects. Examples of those which aim to differentiate the *rate* of educational progress in accordance with the individual differences in the learners' several capacities are the Winnetka and the Dalton plans. Examples of those which aim to differentiate the *amount* or *quality* of learning according to individual differences in the learners' abilities and needs are the Morrison plan and the project or activity programs of instruction. All plans of individual instruction have historical antecedents that reach far into the past. Indeed, it is important to recognize that for centuries individual instruction was the characteristic form and that in America the grade or class organization is less than one hundred years old.

The Winnetka plan. The Winnetka plan is a combination of individual and group instruction. The curriculum is divided into two parts. The first part, dealing with the common essentials of knowledges and skills, is entirely upon an individual basis. The pupil's task is to master the *goals* set up for each grade. In so doing he utilizes an exercise book and appropriate practice materials. When he thinks he has attained a goal, he gives himself a diagnostic test. If the results are unsatisfactory, he continues his practice until he is able to pass another similar test. The final criterion, however, is a mastery test administered by the teacher. As soon as he passes this test, with 100 per cent correct, he goes on to the next unit and proceeds in a similar manner. The amount of work assigned for each

---

[21] *Ibid.*, page 11.

[22] V. V. Caldwell, "Some Facts Regarding Elementary School Trends," *School and Society,* 49: 285–288, March 4, 1939.

TABLE 43

FREQUENCIES WITH WHICH VARIOUS PROVISIONS FOR INDIVIDUAL DIFFERENCES WERE REPORTED IN USE, OR IN USE WITH UNUSUAL SUCCESS, BY 8,594 SECONDARY SCHOOLS IN 1930 (AFTER BILLETT)

NATURE OF PROVISION	PROVISION IN USE		PROVISION IN USE WITH ESTIMATED UNUSUAL SUCCESS		RATIO OF NUMBER OF PROVISIONS IN USE TO NUMBER IN USE WITH ESTIMATED UNUSUAL SUCCESS
	Number	*Per Cent*	*Number*	*Per Cent*	
1. Variations in number of subjects a pupil may carry	6,428	75	795	9	.12
2 Special coaching of slow pupils	5,099	59	781	9	.15
3. Problem method	4,216	49	444	5	.10
4. Differentiated assignments	4,047	47	788	9	.20
5. Advisory program for pupil guidance	3,604	42	540	6	.15
6. Out-of-school projects or studies	3,451	40	439	5	.13
7. Homogeneous or ability grouping	2,740	32	721	8	.26
8. Special classes for pupils who have failed	2,612	30	350	4	.13
9. Laboratory plan of instruction	2,611	30	323	4	.12
10. Long-unit assignments	2,312	27	349	4	.15
11. Project curriculum	2,293	27	365	4	.16
12. Contract plan	2,293	27	465	5	.20
13. Individual instruction	2,145	25	309	4	.14
14. Vocational guidance through exploratory courses	1,911	22	186	2	.10
15. Educational guidance through exploratory courses	1,900	22	193	2	.10
16. Scientific study of problem cases	1,343	16	146	2	.11
17. Psychological studies	1,077	12	70	1	.06
18. Opportunity rooms for slow pupils	946	11	172	2	.18
19. Morrison plan	737	9	175	2	.24
20. Special coaching to enable capable pupils to "skip" a grade or half grade	726	8	114	1	.16
21. Promotions more frequent than each semester	686	8	103	1	.15
22. Remedial classes or rooms	593	7	90	1	.15
23. Adjustment classes or rooms	544	6	55	1	.10
24. Modified Dalton plan	486	6	52	1	.11

TABLE 43 (Continued)

FREQUENCIES WITH WHICH VARIOUS PROVISIONS FOR INDIVIDUAL DIFFERENCES WERE REPORTED IN USE, OR IN USE WITH UNUSUAL SUCCESS, BY 8,594 SECONDARY SCHOOLS IN 1930 (AFTER BILLETT)

NATURE OF PROVISION	PROVISION IN USE		PROVISION IN USE WITH ESTIMATED UNUSUAL SUCCESS		RATIO OF NUMBER OF PROVISIONS IN USE TO NUMBER IN USE WITH ESTIMATED UNUSUAL SUCCESS
	Number	*Per Cent*	*Number*	*Per Cent*	
25. Opportunity rooms for gifted pupils	322	4	69	1	.21
26. Restoration classes	191	2	24	0	.13
27. Dalton plan	162	2	15	0	.09
28. Winnetka technique	119	1	14	0	.12
29. Other techniques	101	1	..	..	..

grade is that which has been found possible of accomplishment within one school year by any normal industrious pupil whose IQ is 95 or above. There are no recitations and no failures; each pupil proceeds at his own rate. About half of each school day is devoted to social and creative activities, which include athletic sports, social studies, fine arts, industrial arts, and dramatics. For these social activities there are no definite academic standards, and an effort is made to group the pupils according to social maturity. While this plan of instruction has been used most widely in the elementary school, it has been employed to some extent in the secondary school as well.

While the Winnetka plan undoubtedly makes better provisions for individual differences than does the conventional school, certain theoretical and practical objections have been made to it. It has been alleged on theoretical grounds that the plan is still curriculum centered rather than child centered, and that there is insufficient integration between the two divisions of the curriculum. Among the practical difficulties that have been suggested are that this organization requires a different type of teacher and of instructional materials from those commonly available, and that it is considerably more expensive. It is probably for these reasons that the plan, although widely discussed, has not been extensively adopted in practice.

The Dalton plan. In 1920 there was tried out at Dalton, Massachusetts, a plan of instruction which was completely individual in

TABLE 44

TRENDS TOWARD GREATER PROVISIONS FOR INDIVIDUAL DIFFERENCES IN ELEMENTARY SCHOOLS (AFTER CALDWELL)

PROVISION	LARGE CITIES		SMALL CITIES	
	Number	*Per Cent*	*Number*	*Per Cent*
A. *Daily Schedule:*				
1. Longer period	23	50	32	55
2. Flexible program	39	85	49	85
3. Subject-matter headings eliminated	17	37	22	38
4. Skills, content, creative activities	35	76	38	66
B *Curriculum Content:*				
1. Child experiences as learning basis	38	83	43	74
2. Elimination of specific subjects	9	20	15	26
3. More freedom for teacher in interpreting course of study	39	85	50	86
4. Emphasis on habits, not fact-learning	34	74	35	60
5. Elimination of drill periods	7	15	13	22
6. Relating learning materials to maturation of child	34	74	37	64
7. Relating learning material to immediate need and mental capacity		85	38	66
8. Experience used to develop number concepts	39 32	70	31	53
9. Delay in formal presentation of abstract arithmetic facts	27	59	37	64
10. Elimination of health as a subject	25	54	33	57
11. Provision for hobby development	39	85	44	76
12. Vacation activity program development	24	52	28	48
C. *Physical Environment:*				
1. Comfortable, adjustable school furniture	32	70	42	72
2. Automatic lighting equipment	11	24	13	22
3. Automatic heating control	30	65	30	52
4. Materials used for sight conservation	21	45	19	33
5. Provision for lunch room	30	65	29	50
6. Provision for rest facilities	28	61	26	45
7. Isolation of sick children	24	52	29	50
8. More floor space per child	17	37	15	26
9. Provision for safe play apparatus	25	54	33	57
10. Provision for ample playgrounds	37	80	41	71
11. Provision for play space in bad weather	18	39	25	43
D. *Materials:*				
1. Basal texts eliminated (skills, content)	10	22	16	28
2. Wide reading material, various levels	40	87	55	95
3. Elimination of work books, etc.	12	21	25	43
4. Variety of material for creative work	40	87	50	86
E. *Classification:*				
1. Provision for pre-school clinics	32	70	38	66
2. Use of reading-readiness tests	38	83	37	64
3. Delay in beginning reading program	29	63	29	50

Table 44 (Continued)

TRENDS TOWARD GREATER PROVISIONS FOR INDIVIDUAL DIFFERENCES IN ELEMENTARY SCHOOLS (after CALDWELL)

Provision	Large Cities		Small Cities	
	Number	*Per Cent*	*Number*	*Per Cent*
4. Groupings by social age, rather than by intelligence or achievement	17	37	15	26
5. Use of no failure program	12	26	16	28
6. Reduction in pupils retained	35	76	41	71
7. Use of tests for guidance, not promotion	37	80	46	79
8. Reduction in number of pupils per teacher	19	41	27	47

character. Since that time it has spread rapidly in England, and somewhat slowly in the United States, generally in modified form. Each classroom becomes a laboratory, with a specialist in charge. There are no recitations in the ordinary sense, although each school day starts with a conference with the teacher. A *contract* consists of the assignments in one subject for a month, and is divided into twenty *units,* one for each day. The pupil accepts in advance a *job,* which includes the assignments in all subjects for one month. The actual amount of time required to do the job depends entirely upon the pupil's learning rate, but he must finish all parts of one job before beginning the next. The plan is not used below the fourth grade. Although pre-tests are given occasionally, and some kind of comprehensive written examination is given at the completion of each job, measurement has a less prominent place than in the Winnetka plan. Much use is made of graphical representation of progress. The plan has been criticized as being distinctly curriculum centered, and as overemphasizing the rugged individualism essential to a pioneer life at the expense of the social co-operation demanded by contemporary civilization.

Simply because all learning is an individual matter in the sense that each person learns from his own activity, it does not follow that all instruction must be individual in character. Nor does it follow that learning cannot take place in, and be facilitated by, the presence of others. As a matter of fact, certain forms of learning can take place in no other way. One learns team work, tact, courtesy, and other similar responses by practice in the social situation. And there is good reason to think that the happiness of the individual and the welfare of society are both dependent more upon an adequate adjustment to the world of people than upon adjustment to the world of things.

The Morrison plan. This plan is based upon the recognition that there are variations in subjects as well as in students. Morrison recognizes five subject types: the science type, the practical-arts type, the language-arts type, the pure-practice type, and the appreciation type. The so-called "Morrison plan" of teaching is mainly applicable to the first two types only. The teaching procedure involves five steps, designated as follows: exploration, presentation, assimilation, organization, and recitation. The material is divided into learning *units,* which are subdivided into *elements,* according to the learner's ability. Usually two to four levels of ability are recognized, with differentiation on the basis of quality as well as quantity. The work for the slower learners is not only less in amount but also easier than that expected of the more capable learners. Testing has a prominent place in the Morrison procedure. Both oral and written tests are sometimes used in the exploratory period, and a mastery test is always used in "the inspection and the acceptance of the completed project." [23] According to Morrison, the mastery test logically comes at the end of the assimilation period of directed study.[24] In actual practice, however, the mastery test often follows the organization or the recitation period, or it may be added as a sixth step in the teaching process. The Morrison "mastery formula" is: "Pre-test, teach, test the result, adapt procedure, teach and test again to the point of actual learning." [25]

It is perhaps worthy of note (see Table 43) that a relatively high percentage of the secondary schools using the Morrison plan in some form thought that it had been an unusual success. In this respect it ranked considerably above the Dalton and Winnetka plans. To meet successfully the individual differences in pupils, however, teachers of unusual ability and training will be required. Given a corps of such teachers, any school will probably be a success, regardless of the plan or scheme of instruction it claims to follow.

The activity movement. In recent years no program of instruction has received more attention among educators than the activity movement, usually a prominent feature of the so-called "progressive schools." Yet educational historians assure us that the principle that man learns by doing is "as old as man's earliest education." [26] In fact, its roots lie further back than the beginning of formal education in schools. Its advocates go even further and assure us that it is grounded in the fundamental nature of the learner himself. However, there are such wide divergencies among its champions, both in theory and in practice, that it may be said that the activity movement not only *recognizes* individual differences to an astonishing degree, but also actually *demonstrates* such differences. The essential features of this educational program may be briefly, and somewhat inadequately, described as follows:

---

[23] Henry C. Morrison, *The Practice of Teaching in the Secondary School,* page 464. Chicago: University of Chicago Press, 1931.

[24] *Ibid.,* page 330.

[25] *Ibid.,* page 81.

[26] Thomas Woody, "Historical Sketch of Activism," in *Thirty-Third Yearbook of the National Society for the Study of Education, Part II,* pages 9–43. Bloomington, Illinois: Public School Publishing Company, 1934. Quoted by permission of the Society.

1. Education results from the child's own purposeful activity with processes considered personally vital to him. An *activity,* according to Kilpatrick, is "a unitary sample of actual child living as nearly complete and natural as school conditions will permit." [27] At every stage the organism reacts as a whole, and the physical, intellectual, and emotional experiences are interrelated.

2. Learning is inherent within the life process itself. It results naturally from the learner's self-directed purposeful activity. Teaching, like learning, is individual in character, arising from a felt need. The teacher is only a guide, and all subject matter is merely a tool. The activity program clearly places upon the shoulders of the classroom teacher the difficult problem of adjustment to individual differences.

3. Interest is at all times the motivating factor in the learning process. Although all teaching procedures recognize the value of interest, the activity movement emphasizes more than any other program the importance of inner drives and interests of the individual pupil, as opposed to extraneous motivation of any kind.

4. The development of the learner's personality, rather than the accumulation of facts and skills, is the objective of all learning. The personality of each individual will develop in accordance with his own abilities, interests, and personal experiences.

5. The evaluation of this relatively intangible personal development involves a fairly long time-span and, therefore, lends itself more to qualitative than to quantitative judgment. In the evaluation process the pupil himself is an active participant. According to Dewey, "the more mature and experienced the teacher, the less will he or she be dependent upon tangible, directly applicable, external tests, and will use them, not as final, but as guides to judgment of the direction in which development is taking place." [28]

It should not be overlooked, however, that regardless of the relative emphasis, such activities as reading and arithmetic are always going to be important, and there appears to be no good reason to rely entirely upon subjective impressions when objective measures are available. The mere fact that adequate measures of the less tangible outcomes are not yet available is no justification for neglecting the measurement of the tangibles, the tools for which do exist. Furthermore, the absence of suitable tools no more removes the need for evaluation than a lack of food relieves the pangs of hunger. Indeed, the need is probably greater, as Gates suggests: [29]

---

[27] *Thirty-Third Yearbook, op. cit.,* page 63. Quoted by permission of the Society.

[28] *Thirty-Third Yearbook, op. cit.,* page 83. Quoted by permission of the Society.

> Any scheme of education that emphasizes the nature and needs of the individual child, as most progressive programs do, has far greater need of measurements than conventional programs designed primarily to impart information and skill to pupils *en masse*.

C. Homogeneous or Ability Groups

Individual and group instruction. It has sometimes been erroneously assumed that there is a necessary conflict between individual and group instruction. While all learning is individual learning, it can take place in a group setting, and certain types of learning can take place only in a group setting; for the individual not only learns *in* the group, he learns *from* the group as well. It has already been pointed out that the two best-known systems of so-called "individual instruction" recognize this fact. It may be debated whether the half day devoted to group activities under the Winnetka plan is sufficient recognition, however, or whether the Dalton plan actually accomplishes pupil co-operation, which it seeks along with freedom and the budgeting of time. In other words, the important question is: What kind of *group organization* best provides for *individual learning?* The problem is to find somewhere between the two extremes of a complete tutorial system and an out-and-out lecture system the program which represents the best possible compromise between that which is educationally ideal and that which is administratively feasible.

Homogeneous or ability groups. Shortly after the development of group intelligence tests in 1917, educational leaders began to use these tests for grouping pupils in school. This procedure was commonly referred to as "homogeneous grouping." It soon became evident, however, that such groups were far from homogeneous, even in intelligence, not to mention other characteristics. The best result that can be obtained under ordinary school conditions is to reduce somewhat the heterogeneity of the instructional groups. The term "ability grouping" came into use as a more accurate term, although frequently used interchangeably with "homogeneous grouping." While much confusion still exists, many writers have recently attempted to make a distinction between these terms. Instructional groups which are made less heterogeneous in learning ability, usually by the employment of general intelligence tests, are called "ability groups." Groups formed upon the basis of some common interest, social maturity, or other similar basis, are called "homogeneous groups." An activity in a progressive school, although made up of pupils of varying abilities, is certainly homogenous from the standpoint of the objective sought. Most of the criticism of grouping is directed against groups formed on the basis of ability. Doubtless, nobody would desire a group possessing the maximum degree of heterogeneity, even in intellectual ability, and certainly not in chronological age, physical maturity, background, motivation, and the like. It is probable, therefore, that *everybody wants a group with a certain degree of homogeneity.* The differences arise regarding the degree and basis of the homogeneity.[30]

---

[29] *Thirty-Third Yearbook, op. cit.*, page 164. Quoted by permission of the Society.

Arguments for and against ability grouping. An imposing list of a dozen or more arguments for, and an equal number against, ability grouping has been assembled.[31] The crucial point at issue is: Do groups formed upon the basis of ability aid or hinder learning? Among the alleged advantages, it is argued that ability grouping makes it easier to adapt instructional materials and methods to the individual pupil, thereby stimulating bright pupils and encouraging dull pupils, with the result that achievement is increased and failure reduced. Among the alleged disadvantages, on the other hand, it is argued that the system is essentially undemocratic and that any gains in academic achievement are likely to be slight in amount and purchased at too dear a price, since the bright pupils tend to graduate too young and to develop a sense of superiority, while dull pupils may overwork or may develop a sense of inferiority. Here as always, however, it is impossible to decide a scientific question merely by counting the arguments pro and con, or by attempting to weigh the logic or fervor with which they are advanced. Fortunately, on this problem a considerable amount of experimental work has been done, although most of the studies must be characterized as inadequate and inconclusive.

The experimental evidence. Adequate summaries of the experimental literature relating to ability grouping have been made by Billett,[32] by Wyndham,[33] by Cornell,[34] and by various writers in the

---

[30] *Cf.* Henry J. Otto, *Elementary School Organization and Administration* (Second Edition), page 184. New York: D. Appleton-Century Company, 1944.

[31] For rather complete summaries of the arguments, see: Austin H. Turney, "The Status of Ability Grouping," *Educational Administration and Supervision,* 17:23, January, 1931; *Ninth Yearbook of the Department of Superintendence,* pages 121–126. Washington, D. C.: National Education Association, 1931; and Ernest W. Tiegs, *op. cit.,* pages 262–264.

[32] Roy O. Billett, *op. cit.,* pages 16–37.

Review of Educational Research.[35] In 1934 a foreign observer[36] commented upon the "haphazard condition" of the research upon the problem and pointed out that the experimental studies "raise more issues than they settle."

Ten years later an American educator[37] could see "little or no solid, objective evidence upon which to base decision as to the effectiveness of homogenous grouping as actually practiced." Cornell states the situation as follows: "Reviewers are generally agreed that the experimental evidence as to the achievement status of pupils under a plan of ability grouping is inconclusive."[38] This writer notes, however, that "one of the most consistent results" has been the increased speed of progress possible by bright learners "at every level from the first grade through college," and that a reduction in the amount of failure by the less capable learners has been "rather consistently reported."[39] Her final conclusion is as follows:[40]

> The results of ability grouping seem to depend less upon the fact of grouping itself than upon the philosophy behind the grouping, the accuracy with which grouping is made for the purposes intended, the differentiations in content, method, and speed, and the technique of the teacher, as well as upon more general environmental influences. Experimental studies have in general been too piecemeal to afford a true evaluation of results, but when attitudes, methods, and curricula are well adapted to further the adjustment of the school to the child, results, both objective and subjective, seem to be favorable to the grouping.

The above statement is worthy of careful study. It seems reasonably clear that the evil effects of ability grouping feared by its opponents need not occur; and, on the other hand, that the alluring advantages claimed by its advocates may not materialize. In other words, there is no money-back guarantee with ability grouping. At best, it merely affords more favorable conditions for doing something about the problem of individual differences. The fundamental adjustments must be in terms of properly differentiated curricula and of teaching methods. On this point Otto says:

---

[33] Harold S. Wyndham, *Ability Grouping,* pages 128–159. Melbourne, Australia: Melbourne University Press, 1934.

[34] Ethel L. Cornell, "Effects of Ability Grouping Determinable from Published Studies," *Thirty-Fifth Yearbook of the National Society for the Study of Education, Part I,* pages 289–304. Bloomington, Illinois: Public School Publishing Company, 1936. Quoted by permission of the Society.

[35] At three-year intervals, beginning with Volume I, 1931.

[36] Harold S. Wyndham, *op. cit.,* page 156.

[37] L. A. Williams, *Secondary Schools for American Youth,* page 290. New York: American Book Company, 1944.

[38] Ethel L. Cornell, *op. cit.,* page 295. Quoted by permission of the Society.

[39] *Ibid.,* pages 396–397. Quoted by permission of the Society.

[40] *Ibid.,* page 304. Quoted by permission of the Society.

All authorities are agreed that no classification scheme can remove the need for adjusting instructional materials and methods to the varying needs of pupils in the group.[41]

Upon certain important issues, unfortunately, there has been little or no experimentation. No one, for example, has determined the effect of various methods of adapting work to pupils of different levels of ability. This is especially important, since the methods actually employed have usually been most effective for dull learners and least effective for bright learners. In most cases, probably, the methods have been those which are used with ordinary heterogeneous groups, and which appear to be least appropriate to the more capable individuals.

Nor has there been any convincing experimental attack to determine the effect of ability grouping upon the work habits and mental health of the pupils. Such meager results as do exist are favorable. Maller [42] found evidence that such desirable social traits as co-operation were developed better under a system of ability grouping. It is a common observation that the best competition in sports, such as golf and tennis, is among those who "play about the same kind of game." Additional evidence that homogeneity is an attribute of natural social groups is afforded by the numerous studies which have shown that there is a positive correlation between friends of all ages, as well as between husbands and wives, on practically all personality traits investigated.[43] Partridge [44] points out that several studies have revealed a greater similarity among friends in mental age than in chronological age. But the main reliance so far has been upon questionnaire studies, of which the most extensive is by Sauvain.[45] One study of the attitude of 645 junior high-school pupils toward ability grouping came to the conclusion that "the great majority are happy and satisfied . . . and that they accept and believe in the grouping that exists as the best situation for them." [46] That the opinions of parents as well as of teachers are favorable to ability grouping in the cities where it is employed is indicated by the following conclusions: [47]

---

[41] Henry J. Otto, *op. cit.*, page 195.

[42] Julius Bernard Maller, *Coöperation and Competition,* page 163. New York: Bureau of Publications, Teachers College, Columbia University, 1929.

[43] Helen M. Richardson, "Studies of Mental Resemblance between Husbands and Wives and between Friends," *Psychological Bulletin,* 36: 104–120, February, 1939.

[44] E. DeAlton Partridge, *Social Psychology of Adolescence,* Chapter V. New York: Prentice-Hall, Inc., 1938.

[45] Walter Howard Sauvain, *A Study of the Opinions of Certain Professional and Non-Professional Groups Regarding Homogeneous or Ability Grouping,* 151 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1934.

[46] Austin H. Turney and M. F. Hyde, "The Attitude of Junior High School Pupils toward Ability Grouping," *School Review,* 39: 606, October, 1931.

[47] Walter Howard Sauvain, *op. cit.*, pages 115, 116.

> On the whole, where grouping is used, parents believe that children are at least as happy, do better work in school, and are correctly sectioned according to ability. . . .
>
> Teachers seem to like ability grouping somewhat more than do the parents. They believe that grouping improves social attitudes, leads to better work by pupils, and increases the happiness of children. . . .

The technique of ability grouping. There is no general agreement as to the best basis for ability grouping. In fact, there is probably no one "best basis." Much depends upon the local conditions, the data available, the nature of the subject, the size of the school, the fundamental philosophy of the school, and the like. It is often true, as one writer suggests, that "the soundest policy in dealing with educational measurements is to obtain objective data and interpret them subjectively."[48] Nor is there uniformity in either theory or practice regarding the number and size of the groups, the proper differentiation in methods and curricula, or the relative emphasis upon acceleration and enrichment for the bright groups.

A useful distinction is made between vertical and horizontal classification. Vertical classification attempts to bring together pupils of approximately the same *status*. The successive grade levels of the ordinary school represent such an attempt. The basis is usually CA, or some combination of CA, MA, and EA. The use of the average of the MA and EA, or the average G-score on an intelligence test and a general achievement test, has much to commend it in the intermediate and upper classes.[49] Horizontal classification means that on any grade level the pupils are further divided according to ability, or *rate* of learning. For this ability grouping in the academic subjects, the IQ, or a combination of IQ and CA, is probably most often employed. Boyer shows how a two-way distribution of IQ and CA, divided by horizontal and vertical lines, may be used effectively for this purpose.[50] In the high school, aptitude tests are sometimes better than general intelligence tests. In other words, the purpose is to bring together for instructional purposes those pupils who represent approximately the same educational and mental status, and who are capable of progressing in the subject at about the same rate. The system should be flexible enough to permit the shifting of pupils from one group to another in any subject whenever it is evident they are improperly classified in that subject. For non-academic subjects, such as woodwork and music, for extracurricular activities, and possibly for the homeroom in high school, the groups may be as heterogeneous as the population of the school. Such a flexible program is inherently democratic. Small schools are of necessity limited to informal groupings made within the classroom.

---

[48] Jacob S. Orleans, *Measurement in Education,* page 286. New York: Thomas Nelson and Sons, 1937.

[49] *Cf.* William M. McCall, *Measurement,* Chapter XI. New York: The Macmillan Company, 1939.

[50] *Thirty-Fifth Yearbook, op. cit.,* pages 199–203. Quoted by permission of the Society.

But the most important problem of adjustment yet remains for the classroom teacher. She must study the individual pupils in her class, whenever necessary must divide them into temporary groups for remedial instruction, and must vary the instructional materials and teaching methods as conditions seem to warrant. In the last analysis, the adjustment of the school to individual differences becomes a teaching problem. As McCall says, "But after all, how pupils are taught and not how they are grouped is the vital matter." [51] Turney puts the matter concisely: [52]

> The actual sectioning is but a minor part of ability grouping; the real job rests with the *teachers*. To adjust subject matter so that a child *can* use his mental ability, and to adjust method so that he *will* use it—these are the outstanding problems, for it is idle to talk of effective development unless children can and do use their mental ability.

Special classes are sometimes formed for pupils at the extremes of the distribution, although in high school those for the very slow learner are about nine times as frequent as those for the very bright.[53] In such classes the teaching is highly individualized. Patience, skill in diagnosing pupil difficulties, and training in mental hygiene are important qualifications for teachers of slow classes. High intelligence, versatility, sound scholarship, and a thorough grounding in psychology are essential qualifications for teachers of special classes for bright pupils.

It is doubtless possible to overdo the idea of "special" classes and schools of one sort or another. Although in a real sense every pupil is unique and should receive special attention, it would certainly be a grave mistake to become so occupied with the "exceptional" pupils as to overlook adequate educational provision for the larger group, who are, to all intents and purposes "perfectly normal." This situation has been satirized as follows: [54]

Johnny Jones has lost a leg,
Fanny's deaf and dumb,
Marie has epileptic fits,
Tom's eyes are on the bum.

---

[51] William A. McCall, *op. cit.*, page 168.

[52] *Thirty-Fifth Yearbook, op. cit.*, pages 113–115. Quoted by permission of the Society.

[53] Roy O. Billett, *op. cit.*, page 196.

[54] Elmer Harrison Wilds, *The Foundations of Modern Education*, page 523. New York: Farrar & Rinehart, Inc., 1942.

Sadie stutters when she talks,
Mabel has T.B.,
Morris is a splendid case
Of imbecility.
Billy Brown's a truant,
And Harold is a thief,
Teddy's parents gave him dope
And so he came to grief.
Gwendoline's a millionaire,
Gerald is a fool:
So every one of these darned kids
Goes to a special school.
They've specially nice teachers,
And special things to wear,
And special time to play in,
And a special kind of air.
They've special lunches right in school.
While I—it makes me wild!—
I haven't any specialties,
I'm just a normal child.

Acceleration and retardation. In the elementary school a common device for reducing the heterogeneity of the class is to eliminate the extremes of the distribution at promotion time. To do this a small number of the most capable pupils are allowed to "skip" a grade or half grade, and usually a larger number of the least capable pupils are "failed," or "retained" in the same grade for another year or half year. Witty and Wilkins [55] published a critical survey of the literature relating to acceleration, and, in spite of certain limitations in the studies, concluded that "most reports show clearly that acceleration, when practiced, is associated with desirable adjustment in *all* types of development for which data have been assembled." One of the few controlled experiments so far reported, in which pupils allowed to skip a grade were paired with pupils of like ability not skipped, concluded that "under reasonably favorable conditions skipping is a satisfactory method of accelerating pupils of superior ability." [56]

Recent studies have attempted to determine the effect of acceleration upon the pupils' personality and social adjustments in high school and college, apparently accepting Terman's verdict regarding the academic achievement of superior pupils: "The earlier they enter college the better work they do there, at least down to an entrance age of 15 years."[57] Almost without exception the results appear to be favorable. Engle, for example, found that accelerated students in high school when compared with other students of their own chronological age were "at least as active socially as non-accelerated students."[58] In 1943, Pressey[59] surveyed the literature regarding acceleration on the college level and came to the conclusion that "the great majority of accelerated students do well in school, are socially adjusted, do not suffer in health, and are not handicapped in after-school career."

---

[55] Paul A. Witty and Laroy W. Wilkins, "The Status of Acceleration or Grade Skipping as an Administrative Practice," *Educational Administration and Supervision,* 19: 321–346, May, 1933.

[56] Jesse E. Adams and C. C. Ross, "Is Skipping Grades a Satisfactory Method of Acceleration?" *American School Board Journal,* 85: 24–25, July, 1932.

During World War II great emphasis was placed upon accelerated programs of education, particularly on the college level. Several colleges have attempted to investigate the effect of these programs upon the students. Studies directed by Pressey[60] at Ohio State University have been especially noteworthy. With few exceptions the results have favored acceleration. Another recent investigation[61] concludes that many more superior women students than usually attempt it "can complete a college program in three years or less without unfortunate effects as regards scholarship, recreation, health, or after-school career."

The weight, both of the arguments and of experimental evidence, appears to be against failure or retardation as a school policy. In Otto's[62] survey the literature indicated that about 20 per cent of repeaters do better and 40 per cent do worse than before. He concluded that if the objective of the modern school is the optimum development of its pupils, "non-promotion is not the way to get it." Several studies have been reported during the past twenty-five years which indicate the value of trial promotions. An investigation by McKinney,[63] for example, involving more than 13,000 pupils, shows a saving of about three out of every four repeaters. One study has shown that the threat of failure affords ineffective motivation.[64] In the words of an acknowledged authority in the field, both the logic and the evidence in the case "point to the unquestioned conclusion that 'school failure,' that is, the repetition of grades, should be abandoned as an administrative device." Certainly, with a modern curriculum and an adequate program of diagnosis and guidance, few if any failures should occur.

---

[57] Lewis M. Terman, "The Gifted Student and His Academic Environment," *School and Society,* 49: 68, January 21, 1939.

[58] Thelburn L. Engle, "A Study of the Effects of School Acceleration upon the Personality and Social Adjustments of High-School and University Students," *Journal of Educational Psychology,* 29: 523–529, October, 1938.

[59] S. L. Pressey, "Acceleration versus Lock Step," *Educational Research Bulletin,* 22: 29–35, February 17, 1943.

[60] *Cf.* S. L. Pressey and S. B. Folk, "First Evaluations of an Accelerated Program in a College of Engineering," *Journal of Engineering Education,* 34: 477–485, March, 1944; S. L. Pressey, "Acceleration the Hard Way," *Journal of Educational Research,* 37: 561–570, April, 1944.

[61] Marie A. Flesher, "An Intensive Study of Seventy-Six Women Who Obtained Their Undergraduate Degrees in Three Years or Less," *Journal of Educational Research,* 39: 602–612, April, 1946.

[62] Henry J. Otto, *op. cit.,* page 232.

[63] H. T. McKinney, *Promotion of Pupils, A Problem in Educational Administration,* 206 pages. Urbana, Illinois: University of Illinois, 1921.

Continuous promotion. Otto [65] has proposed a somewhat theoretical but very suggestive promotion plan for the elementary school, which abolishes not only acceleration and nonpromotion but the term "school grade" as well. Such a type of organization has been in successful operation in several school systems for a number of years.

His plan involves the following five essential features:

1. There would be available extensive data of an objective character on each child, so that he may be "placed at all times in groups in which he can work to the best advantage in terms of his own developmental readiness."
2. There would be continuous pupil adjustment and progress with shifts from one group to another "at any time during the year that a change would seem advisable."
3. The major classifications which take place in the ordinary school at the beginning of each term would be eliminated.
4. It would make possible longer teacher-group relationships in which "the same teacher works with the same group of children for two or three consecutive semesters or years."
5. The conventional competitive marking system would be replaced with "extensive, objective, cumulative data on many aspects of the growth and development of each child."

SELECTED REFERENCES FOR FURTHER READING

Billett, Roy O., *Provisions for Individual Differences, Marking and Promotion.* Washington, D. C.: United States Office of Education, 1933. Parts I, II, and III.

Cox, Warren W., and others, "The Grouping of Pupils," *Thirty-Fifth Yearbook of the National Society for the Study of Education, Part I.* Bloomington, Illinois: Public School Publishing Company, 1936. 315 pages.

Elsbree, Willard S., *Pupil Progress in the Elementary School.* New York: Bureau of Publications, Teachers College, Columbia University, 1943. 86 pages.

Gilliland, A. R., and Clark, E. L., *Psychology of Individual Differences.* New York: Prentice-Hall, Inc., 1939. 535 pages.

---

[64] Henry J. Otto and Ernest O. Melby, "An Attempt to Evaluate the Threat of Failure as a Factor in Achievement," *Elementary School Journal,* 35: 588–596, April, 1935.

[65] Henry J. Otto, *op. cit.,* 236–242.

Jacobson, Paul B., and Reavis, William C., *Duties of School Principals.* New York: Prentice-Hall, Inc., 1941. Chapters XII and XIII.

Lee, J. Murray, *A Guide to Measurement in Secondary Schools.* New York: D. Appleton-Century Company, 1936. Chapter VI.

McCall, William A., *Measurement.* New York: The Macmillan Company, 1939. Book Three.

Morrison, Henry C., *The Practice of Teaching in the Secondary School.* Chicago: University of Chicago Press, 1931. 688 pages.

Odell, C. W., *Educational Measurement in High School.* New York: D. Appleton-Century Company, 1930. Chapter XXI.

Orleans, Jacob S., *Measurement in Education.* New York: Thomas Nelson and Sons, 1937. Chapter IX.

Otto, Henry J., *Elementary School Organization and Administration* (Second Edition), New York: D. Appleton-Century Company, 1944. Chapters IV–VII, XII.

Symonds, Percival M., *Measurement in Secondary Education.* New York: The Macmillan Company, 1927. Chapters XXII and XXIII.

Tiegs, Ernest W., *Tests and Measurements in the Improvement of Learning.* Boston: Houghton Mifflin Company, 1939. Chapter XI.

Umstattd, J. G., *Secondary School Teaching* (New Edition). Boston: Ginn and Company, 1944. Division II.

CHAPTER XVI

Guidance

A. The Problem of Guidance

The meaning of guidance. The fundamental problem of life is adjustment. At birth the human infant is much less well adjusted to the world in which he must live than many of the simpler organisms. Man's dominant place in the universe is due largely to his remarkable capacity for modifying his reactions in the direction of a more adequate adaptation to the conditions under which he must live. The process by which these changes take place is called *learning,* and the result is called *education.* The function of the school is to provide a favorable environment in which these changes may take place. The role of the classroom teachers and of the school administrators is to stimulate and to direct the learning process. This directive function is called *guidance.*

The aim of all guidance is to assist the learner to acquire sufficient understanding of himself and of his environment to be able to utilize most intelligently the educational opportunities afforded by the school and the community. The problem of guidance arises from the fact that an immature but growing individual with a unique combination of abilities and limitations is confronted with a complex and ever-changing environment. Guidance used to be regarded as an effort "to see through Johnny and to see Johnny through." The emphasis today has shifted to an effort "to help Johnny see through himself and to see himself through." [1] It seeks to assist each student to choose, and make satisfactory progress in, those activities which will contribute most to his development, individual happiness, and social worth.

The importance of guidance. Certain circumstances have conspired to make guidance one of the most acute problems of the modern school. This is particularly true of the secondary school and of the college.[2] In the four decades from 1890 to 1930, the total population of the United States doubled, but the enrollments in the higher institutions of learning increased approximately tenfold, and in the secondary school twentyfold. Growth since 1930 has been less rapid, but in round numbers, since 1890 the population of the secondary school has doubled every ten years, and that of the institutions of higher learning every fifteen years. As a result of these conditions, the student body of the modern secondary school and college represents a greater diversity of backgrounds, interests, ambitions, and abilities than has ever been true before.

---

[1] George E. Myers, *Principles and Techniques of Vocational Guidance,* page 4. New York: McGraw-Hill Book Company, 1941.

[2] Leonard V. Koos and Grayson N. Kefauver, *Guidance in Secondary Schools,* Chapter I. New York: The Macmillan Company, 1932.

At the same time science and invention have greatly complicated and are constantly changing the social and economic world from which these pupils come and to which they must return. Likewise, the school situation itself, academically as well as socially, has greatly increased in complexity. The small high school with a single curriculum leading to college has tended to give way to larger schools with a more diversified program. Judd[3] has called attention to the fact that the number of subjects offered in American high schools increased from 9 in 1890 to more than 250 in 1942. At the present time a pupil of high-school age in a modern American city has a choice of a score or more different curricula.

As it is always easier for the traveler to lose his way in a large city than in a small town, especially if he is lacking in maturity and experience, it is perhaps not surprising that the majority of those who enter the modern secondary school and college never succeed in making a satisfactory adjustment to these institutions. Likewise, the vast number of adolescents and adults who find their way into penal institutions or into hospitals for the physically and the mentally ill, and the much larger number of others who lead unhappy and unsuccessful lives afford tragic evidence that the adjustment outside the school has been equally unsatisfactory. There seems no escaping the fact that when the conditions of life increase in complexity, the need for guidance increases proportionately. The better the guidance program the less will be the need for diagnostic and remedial work later on. An adequate guidance program is the best form of prevention.

It is significant that the word *guide* or *direct* appears in five of the ten functions recognized by the Committee on the Orientation of Secondary Education, Department of Secondary School Principals, National Education Association.[4] The guidance function is also clearly implied in most, if not all, of the other five. A few years ago a prominent American educator[5] asserted that guidance "has been the greatest single force in the improvement of education in this country since the pioneer attempt early in the present century to apply quantitative measures to school processes."

---

[3] Charles H. Judd, "General Education and the Baccalaureate Degree," *School and Society,* 56: 35, July 11, 1942.

[4] *Bulletin of the Department of Secondary School Principals of the National Education Association,* Volume 21, Number 64, 266 pages. Washington, D. C.: National Education Association, January, 1937.

In spite of this clearly recognized need for guidance in the modern secondary school, there is a considerable amount of evidence that the need is not being adequately met at the present time. Symonds traces six lines of professional development which now converge upon guidance, and describes the resulting guidance programs in the public schools as "indescribably chaotic."[6] The committees who visited 200 secondary schools reported in 1938 that "the guidance service is probably less well organized and is operating less effectively than any other phase of secondary school activity."[7] Approximately two fifths of the 17,000 pupils in these schools stated that the value of the guidance to them was either very little or none at all.[8] After a careful summary of the attempts that appeared from 1932 to 1937 to evaluate vocational guidance, Kitson and Crane concluded that the evidence advanced was "pitifully insignificant when compared with the momentous aims of vocational guidance."[9]

The scope of guidance has greatly increased in recent years. Guidance began in America when standard tests of achievement and of intelligence appeared as vocational guidance in the hands of special workers. The scope has been extended, until today guidance is coming to be regarded as an "inseparable aspect of the educational process."[10]

This enlarged scope of guidance was an inevitable accompaniment of the enlarged concept of education in the modern school. The aim of education is no longer considered merely the mastery of a static curriculum largely informational in character, but is now regarded as the development of the whole child. Therefore problems arise in the life of every pupil rather than become the monopoly of a small group of so-called "problem pupils." For example, at the University of Minnesota, where an extensive guidance program has been carried on, individuals in the highest tenth of ability have fully as many problems as those in the lowest tenth.[11] While it may still be true, as Allen[12] contends, that guidance is "primarily the principal's responsibility," the sheer magnitude of the task requires the active co-operation of everybody connected with the school, although in the larger schools the general administration and the more technical aspects of the program may be placed in the hands of certain specialists in guidance.

---

[5] Harold Benjamin, "Editor's Introduction" to *Principles and Techniques of Vocational Guidance* by George E. Myers, *op. cit.*, page xii.

[6] Percival M. Symonds, "A Plea for the Integration of School Guidance Activities." *Teachers College Record,* 38: 686–710, May, 1937.

[7] M. L. Altstetter, "Guidance Service in Two Hundred Secondary Schools," *Occupations,* 16: 513–520, March, 1938.

[8] Summarized from an address by Walter C. Eells, *Occupations,* 16: 760, May, 1938.

[9] Harry D. Kitson and Margaret Crane, "Measuring Results of Vocational Guidance, A Summary of Attempts, 1932–1937," *Occupations,* 16: 837–842, June, 1938.

[10] *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I,* page 24. Bloomington. Illinois: Public School Publishing Company, 1938. Quoted by permission of the Society.

While guidance should be regarded as a unified and continuous process, at least three phases of the process may be differentiated, according to the major emphasis given: namely, vocational, educational, and personal guidance. The service which such guidance should render the individual has been stated as follows:[13]

> In the first place, it should enable him to obtain as objective and as clear a picture of himself as modern scientific techniques and the ingenuity of educators, counselors and special advisers are able to portray. Secondly, it should make known to him the opportunities, educational, vocational and social, which are at hand in the school environment and those existing beyond its doors. In the third place, it should attempt to guide him toward those opportunities which are available and appropriate to his particular needs and capacities.

The place of measurement in guidance. Two errors are common in assigning the place of measurement in guidance. The first of these, fortunately now less common than in the early days of standard testing, is to think of guidance as synonymous with testing. *Guidance is always more than the giving of tests, no matter how extensively or carefully done.* As a matter of fact, whether or not tests serve any guidance function depends upon the use made of the results. Here, as elsewhere, tests are merely tools. The second error, unfortunately very common today, is to dismiss measurement altogether and to regard it as wholly unessential to guidance if not indeed an actual obstacle. Rogers,[14] for example, a leading exponent of so-called "client-centered counseling," holds that the process is more likely to be successful if the major responsibility is at all times centered in the client himself rather than in the counselor, who is making a sincere attempt to accept the client as he views himself. This objective can be attained best when the counselor drops all efforts to evaluate and diagnose, asks no questions and volunteers no advice. Rogers contends that the use of tests tends to center the responsibility in the counselor rather than in the client and to emphasize external evaluating rather than self-appraisal, which he regards as far more important.

---

[11] E. G. Williamson and J. G. Darley, *Student Personnel Work,* page 258. New York: McGraw-Hill Book Company, Inc., 1937.

[12] Richard D. Allen, "How a Principal Can Direct Guidance," *Occupations,* 16: 15–20, October, 1937.

[13] Donald G. Paterson, Gwendolen G. Schneidler, and Edmund G. Williamson, *Student Guidance Techniques,* pages 28–29. New York: McGraw-Hill Book Company, Inc., 1938.

[14] Carl R. Rogers, "Psychometric Tests and Client-Centered Counseling," *Educational and Psychological Measurement,* 6: 139–144, Spring, 1946. For a fuller statement, see: Carl R. Rogers, *Counseling and Psychotherapy.* Boston: Houghton Mifflin Company, 1942.

1. APPRAISAL By means of
    - a. Tests.
    - b. Rating scales.
    - c. Diagnostic records.
    - d. Autobiography.
    - e. Interviews.
    - f. Observations, anecdotes, etc.
    - g. Group discussion.
    - h. Psychiatric case study.
    - i. Case history, cumulative records.
2. ADJUSTMENT By means of
    - a. Changes in the environment through
        - 1. Curriculum.
        - 2. Extra-class activities.
        - 3. Physical environment.
        - 4. Discovery and provision of educational and vocational opportunities.
    - b. Changes in the individual through
        - 1. Interviews.
        - 2. Lectures.
        - 3. Group discussions.

Figure 48. The Complete Scope of Guidance. (After Strang.)

This viewpoint is as extreme as the first however. While testing is never everything in guidance, it is always something. In fact, it may be confidently asserted that *evaluation in some form is implicit in the guidance function.* Properly used tests are indispensable in self-analysis. Two aspects of guidance are often distinguished, appraisal and adjustment. It is clear from Figure 48, adapted from Strang's[15] analysis, that the second aspect is dependent upon the first, and that the first has to do with measurement or evaluation in some form. As Bingham[16] puts it, testing is "one of the indispensable handmaids of guidance."

Fowler[17] indicates that many common mistakes will be avoided if users of tests in guidance will remember at all times that "the only justifiable reason for using tests in the guidance program is to serve the individual inventory in counseling." He formulates seven "guiding rules" based upon this point of view:

---

[15] Ruth Strang, *The Role of the Teacher in Personnel Work,* page 14. New York: Bureau of Publications, Teachers College, Columbia University, 1935.

[16] Walter V. Bingham, "A National Perspective on Testing and Guidance," *Educational Record,* 20: 137–150, Supplement No. 12, January, 1939.

[17] "To Inquirers about Tests," *Education for Victory,* 3: 12–13, December 4, 1944.

1. Any item of the individual inventory, whether it be a test score, a teacher's mark, a fact about the pupil's health, can be interpreted in the counseling situation only in the light of all the other inventory data having some bearing on the problem at hand. This is to say, a chief value of test scores is the check which they provide upon the meaning of other accumulated facts. In turn, the importance to be accorded test scores in any given case must be weighed in the light of other data from the individual inventory. Dependence must be placed upon tests to supply facts when they have not been accumulated through other means.
2. Test scores, like other items in the inventory, must be interpreted cautiously until norms are scientifically established for the local situation and for the particular kind of problem which the pupil presents.
3. The meaning of a test score may not be the same from one pupil to another because of the differences in other pertinent inventory data. The meaning may change even for the same pupil from one problem to another or from one time to another.
4. Real counseling will encourage decisions or judgments only on the basis of as full an inventory of pertinent facts as possible. Thus several measures are usually better than just one or two. Likewise, the same dependence will not be placed upon so-called "interest" or "personality" tests as upon achievement and aptitude tests.
5. It is recognized that certain tests are regularly used *in the school* by the administrator in pupil classification and curriculum planning. They are used by teachers in individualizing teaching methods. The data from these same tests are of even greater use for counseling and should always be recorded in the cumulative record. Tests used by the administrator for these purposes may supplement the tests used only by the counselor. This fact should not be overlooked in their choosing.
6. Tests are best used as aids to *counseling*, rather than as standards for arbitrary selection (or rejection) for training and job opportunities.
7. Familiarity with a test, gained through its use, is important. In deciding to use a new test to measure the same traits, loss of this familiarity should be weighed carefully against the possible gain in reliability, validity, usability, and economy.

If these suggestions are kept in mind, the dangers against which Rogers warns will be avoided. Now let us consider more fully the role of measurement in a sound program of guidance.

B. The General Technique of Guidance

Life has often been described as a journey to a far country. The particular destination selected by the traveler will depend upon his interests, the information concerning the advantages offered there, and the resources available for the journey. The wise traveler plans the journey with care; he reads the road maps and other guides, and if possible consults experienced persons who have been over the road. Only in this way may he reasonably hope to avoid needless detours and to take advantage of the short cuts and scenic routes. It is not unlikely, however, that, no matter how carefully the trip is planned in advance, circumstances will arise along the way that will make continuous revisions necessary. In the words of a recent book: [18]

> Tests and guidance are merely aids to travel. They are the signposts and the road maps which point the way and eliminate miles of unsure exploration.

Kitson gives the following concise statement of the technique of vocational guidance: [19]

> Thus understood vocational guidance comprises a variety of services: (1) analysis of occupations, which will lead to exact information regarding the requirements of each one, the conditions of work obtaining, and the rewards that may be expected . . . ; (2) the analysis of the individual, through which may be discovered the degree to which he can meet the requirements, and the degree to which he can conform to the conditions of work in any occupation he may be considering; (3) counsel and advice regarding the solution of occupational problems, not merely in school, but after one has completed his formal education; (4) organized placement through which individuals can obtain free assistance in finding jobs; and (5) follow-up service of information, counsel, and replacement.

While the fourth step as stated would appear to be primarily applicable to vocational guidance, the others are equally essential to educational and personal guidance. Each of these steps will now be discussed briefly.

Analysis of opportunities available. It is obvious that before an individual can make a satisfactory adjustment to the situation, he must recognize its opportunities and limitations. Authoritative statements, given either orally or in written form, are essential. In educational guidance, information regarding the history, traditions, purposes, and general organization of the school is needed. This may be provided by lectures to groups, by discussions with individuals and by publications, such as bulletins, handbooks, and programs of studies. Definite courses on occupations and explanatory or try-out courses are frequently available for vocational information, as well as suggested library reading, motion pictures, and opportunities for actual visitation and observation of various types of work being done. In personal guidance the resources most important for the development of the individual personality may be the extracurricular activities of the school and the community.

---

[18] Donald G. Paterson, Gwendolen G. Schneidler, and Edmund G. Williamson, *op. cit.*, page 29.

[19] Harry Dexter Kitson, "Distribution of Workers among the Occupations," *Teachers College Record*, 34: 465-466, March, 1933.

Analysis of the individual. Before an individual is ready to make an intelligent choice of an educational program in school or of a vocation in life, he must have dependable knowledge of his own strong and weak points. "Guidance," as Ruch and Segel said, "tend to be effective to the degree that we can draw up a balance sheet for each individual, upon which we can record in objective terms the strengths and weaknesses, the peaks and valleys, of his physical, mental, and social capacities." [20] No matter how complete and accurate the information regarding the prospective vocation that may have been secured from observation trips, interviews with successful persons, explanatory and try-out courses, and the like, it is insufficient. The best such information can do is to give the individual some idea of whether or not he will probably *like* the type of work and whether he *thinks* he can do it. The unreliability of self-analysis, however, has been well established by many studies since the pioneer work of Hollingworth.[21] Fortunately, considerable progress has been made in recent years in developing valuable information about the person, which he needs in making his choice, and which he can hardly be expected to find out about himself unaided. Most of these appraisal procedures are listed in Figure 48.

Traxler [22] has provided descriptions of available tests with practical suggestions as to their use in guidance. He recommends that the basic program in high school should consist of at least five tests annually: "a test of academic aptitude or reading on alternate years, and achievement tests in English and three of several other fields, such as mathematics, science, social studies, foreign languages, commercial subjects, fine arts, and practical arts—depending on what the pupil is studying." [23]

It will be noted that these techniques of evaluating the individual personality are of three types. Some are based upon the direct observation of the individual. Tests are merely one form of controlled observation. Behavioral descriptions in the form of anecdotal records, when skillfully done, are of considerable value.

---

[20] Giles M. Ruch and David Segel, *Minimum Essentials of the Individual Inventory in Guidance,* page v. Washington: United States Office of Education, 1939. This monograph of 83 pages is the most satisfactory concise treatment of various aspects of the problem available today.

[21] H. L. Hollingworth, *Judging Human Character,* Chapter IV. New York: D. Appleton and Company, 1922.

[22] Arthur E. Traxler, *Techniques of Guidance,* Chapters III–X. New York: Harper & Brothers, 1945.

[23] *Ibid.,* page 11. For a brief discussion, see *Educational and Psychological Measurement,* 6: 3–16, Spring, 1946. For a valuable, concise discussion of the preparation and use of these records, see Arthur E. Traxler, *op. cit.,* Chapter VII. For a thorough discussion of the whole problem see Daniel A. Prescott and Staff, *Helping Teachers Understand Children,* 468 pages. Washington: American Council on Education, 1945.

Other techniques involve questioning the individual. This may be done by holding a personal interview, by having him write his autobiography, or by using a written questionnaire. A tactfully handled interview in which the pupil is encouraged to do most of the talking is likely to be especially revealing. Still other techniques rely upon asking other persons about the individual. Rating scales are good examples of this approach. All of these methods are useful, and all have their limitations. Rarely is one of them alone sufficient for an adequate appraisal.

While these diagnostic tools should always be selected and applied by one trained to use them skillfully, Allen [24] has called attention to the possibilities of making the pupil a "partner in the project." He points out that [25] "an understanding of the implications of the test may be the best possible kind of guidance, especially if it leads to more accurate self-appraisal, necessary remedial measures, increased incentive for achievement, and wiser choices of educational and vocational opportunities." Troyer and Pace [26] take a similar position:

> The processes of evaluation are more likely to be in harmony with a democratic philosophy of human relations when the major responsibility for appraisal is carried by the learner than when it is placed on a teacher or expert. When the responsibility is carried by the specialist there is a tendency to administer tests *en masse* for convenience, to interpret the results for the students, to tell them in what ways they are strong and weak, and often to prescribe what steps they should take next.

Eurich and Wrenn list the "chief kinds of information that must be studied in order that counselor or teacher may understand the pupil and the pupil understand himself" as follows: [27]

1. The record of his previous school experience.
2. His aptitudes and abilities.
3. His home background and community environment.
4. His goals and purposes.
5. His interests, likes, and dislikes.
6. His social development and adjustment.
7. His emotional status.
8. His health record and present health status.
9. His economic and financial status.

---

[24] Richard D. Allen, *Self-Measurement Projects in Group Guidance,* pages 3–16. New York: Inor Publishing Company, 1934.

[25] *Ibid.,* page 13. Italicized in the original.

[26] Maurice E. Troyer and C. Robert Pace, *Evaluation in Teacher Education,* page 88. Washington: American Council on Education, 1944.

[27] *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I, op. cit.,* pages 33–44. Quoted by permission of the Society.

Expertness is especially required in the *interpretation* of such information. What is needed is a total picture of the individual, past developmental history as well as present status. This synthesis is more difficult than analysis and is dependent upon it. The guidance value of very high and very low scores is especially great. This is true both of the peaks and valleys of the individual's profile and of the extreme scores in the total distribution. It must be recognized that comparisons are meaningless unless based upon norms derived from similar groups.

In a recent practical discussion Bixler and Bixler [28] stress the point that counselors should avoid persuasive methods and should remain strictly neutral toward the test and the client's reactions to these data, at the same time taking necessary action to facilitate the client's self-evaluation and subsequent decisions by the use of appropriate therapeutic procedures. Briefly, the responsibility of the counselor is "to give the client information, clarify his attitudes toward that information and towards his limitations, and finally to assist him in implementing his plans."

Counseling. Counseling is the process of assisting the individual in making the maximum adjustment to the educational opportunities of his environment in terms of his abilities, interests, and needs. It is normally a face-to-face relation between an older, more experienced person and a less mature person. It is an example of cooperative problem solving. The role of the counselor is not to make decisions for the pupil, but rather to help him intelligently to solve his own problems. The pupil's part naturally increases with his maturity, the ultimate purpose of counseling being so to develop the pupil's self-reliance that outside help becomes progressively unnecessary.

Counseling is not a new development in education. In some type or other it has probably existed longer than formal education itself. The difficulty has been not with the amount of counseling available but with its quality. Nowhere is the wisdom of the homely American philosopher, Josh Billings, more apparent than in counseling: "It is better to kno less, than to kno so mutch that ain't so." The improvement of measurement techniques in the present century and the development of cumulative record forms have made it possible to substitute factual data for opinion and hearsay. However, after studying the uses made of test results in guidance in 493 secondary schools, Lee [29] came to the conclusion that probably fewer than 10 per cent of these schools were making effective use of tests for this purpose, and that this usage was largely restricted to intelligence tests mainly employed as an aid in studying and advising failing pupils.

---

[28] Ray H. Bixler and Virginia H. Bixler, "Test Interpretation in Vocational Counseling," *Educational and Psychological Measurement,* 6: 145–155, Spring, 1946.

[29] J. Murray Lee, *A Guide to Measurement in Secondary Schools,* pages 132–133. New York: D. Appleton-Century Company, 1936.

An excellent example of the type of information needed for intelligent counseling is given in Figure 49, adapted from Fryer and Sparling.[30] It will be noted that these authors have suggested the general achievement levels to be expected of persons whose adult intelligence levels are given. They have also interpreted these levels in terms of both educational and vocational performance. Figure 50 shows the great overlapping of all occupational groups, however. It is, of course, possible to misuse such information, for other data besides intelligence must always be taken into account. Furthermore, the level of educational and vocational achievement an individual attains is dependent to a considerable extent upon the local situation, such as the competition afforded and the standards in operation. It is doubtful if ready-made techniques can ever be worked out which will be equally applicable to all individuals and to all situations. For some time to come most of the critical decisions in life will depend upon a careful consideration of the pertinent data available, in the interpretation of which expert advice and counsel are essential. And, as is usual where human judgment is involved, the consensus of a group of competent persons is likely to be better than the verdict of a single individual. This is one of the greatest values of an educational clinic.

Before any factor can safely be employed as a basis for guiding pupils, previous experimental results must have shown it to be associated with achievement in a certain line. The coefficient of correlation is the principal technique for determining the degree of relationship that exists. Usually the maximum prediction can be made by combining two or more factors at their optimum weights into a single regression equation. The detailed discussion of this somewhat technical procedure is beyond the scope of the present volume, which is designed primarily for consumers of educational research rather than for producers.[31]

It is important, however, that anyone who attempts to counsel pupils know how to utilize the basic research work that must underlie scientific guidance. Fortunately, this can be done in a fairly nontechnical manner. The most important fact to be kept in mind is that all predictions are in terms of probability. The best that

---

[30] Douglas Fryer and E. J. Sparling, "Intelligence and Occupational Adjustment," *Occupations*, 12: 55–63, June, 1934.

[31] A brief discussion is found in Chapter VIII, and a more adequate treatment can be found in any of the books on educational statistics listed at the end of the chapter.

INTELLIGENCE GROUP	ACHIEVEMENT LEVELS		
	General	*Educational*	*Occupational (Examples)*
A. Very superior 18.0 and up Mental Age	Intelligence for creative and directive effort. *High professional occupational level.*	Ability for superior (honor) record in university.	Editor, lawyer, college and high school teacher, engineer, diplomat, minister, business executive, etc.
B. Superior 16.5 to 17.9 Mental Age	Intelligence for executive business, leadership, and most professional endeavor. *Professional occupational level.*	Ability for an average college record.	Journalist, physician, elementary school teacher, large merchant and banker, chemist, social worker, dentist, private secretary, etc.
C+. High Average 15.0 to 16.4 Mental Age	Intelligence for *minor* executive and leadership positions. Excellent capacity for abstract detailed and highly skilled mechanical work. *Technical occupational level.*	Ability for secondary school graduation and some college training.	Stenographer, bookkeeper, nurse, office clerk, teacher of special subjects, photographer, telegrapher, musician (band), radio operator, etc.
C. Average 13.0 to 14.9 Mental Age	Intelligence for routine and skilled mechanical work. Rarely capable of complicated abstract, detailed work. *Skilled occupational level.*	Ability for elementary school graduation and some secondary school training.	Locomotive engineer, telephone operator, policeman, auto mechanic, plumber, chauffeur, tailor, farmer, barber, bricklayer, etc.

Figure 49. Achievement Levels Corresponding to Various Levels of Adult Intelligence. (After Fryer and Sparling.)

can be done is to state the chances of success for any given score. The most useful probability tables are those based upon the experience of the particular school. All that is required is to employ some test or combination of tests that have been found by research workers to be correlated with achievement along the line in which a prediction is sought and then to tabulate the results in appropriate

INTELLIGENCE GROUP	ACHIEVEMENT LEVELS		
	General	*Educational*	*Occupational (Examples)*
C—. Low Average 11.0 to 12.9 Mental Age	Intelligence for some skilled routine work. *Semiskilled and low-skilled occupational level.*	Ability rarely sufficient for elementary school graduation.	Hospital attendant, mason, lumberman, watchman, shoemaker, sailor, leather worker, porter, laborer, etc.
D. Inferior 9.5 to 10.9 Mental Age	Intelligence for *simple* work only. Requires unusual amount of supervision. *Unskilled occupational level.*	Ability so limited that individual usually drops out of elementary school before fifth grade.	Fisherman, lifter, unskilled laborer, loader.
D—. Very inferior 7.0 to 9.4 Mental Age	Intelligence for *very* simple routine work only. Lacks self-direction entirely. *Lowest unskilled occupational level.*	Ability so limited that individual is rarely capable of advancement beyond third grade elementary school.	Laborer (simplest work).
E. Useless 0.0 to 6.9 Mental Age	Intelligence for no social effort. Sometimes possible for high-grade imbeciles to do *very simple* routine tasks under very careful supervision.	Ability so limited that individual is rarely capable of making any advancement in elementary school. May be able to do normal work usually offered in kindergarten.	No occupation.

Figure 49. Achievement Levels (*Continued*).

tables. Such tables are very useful guides in future counseling.

Table 45 illustrates a simple table which shows the probability of success and failure in geometry corresponding to various scores on the Lee Test of Geometric Aptitude.[32] The last column shows

---

[32] J. Murray Lee and Dorris May Lee, "The Construction and Validation of a Test of Geometric Aptitude," *Mathematics Teacher,* 25: 199, April, 1932.

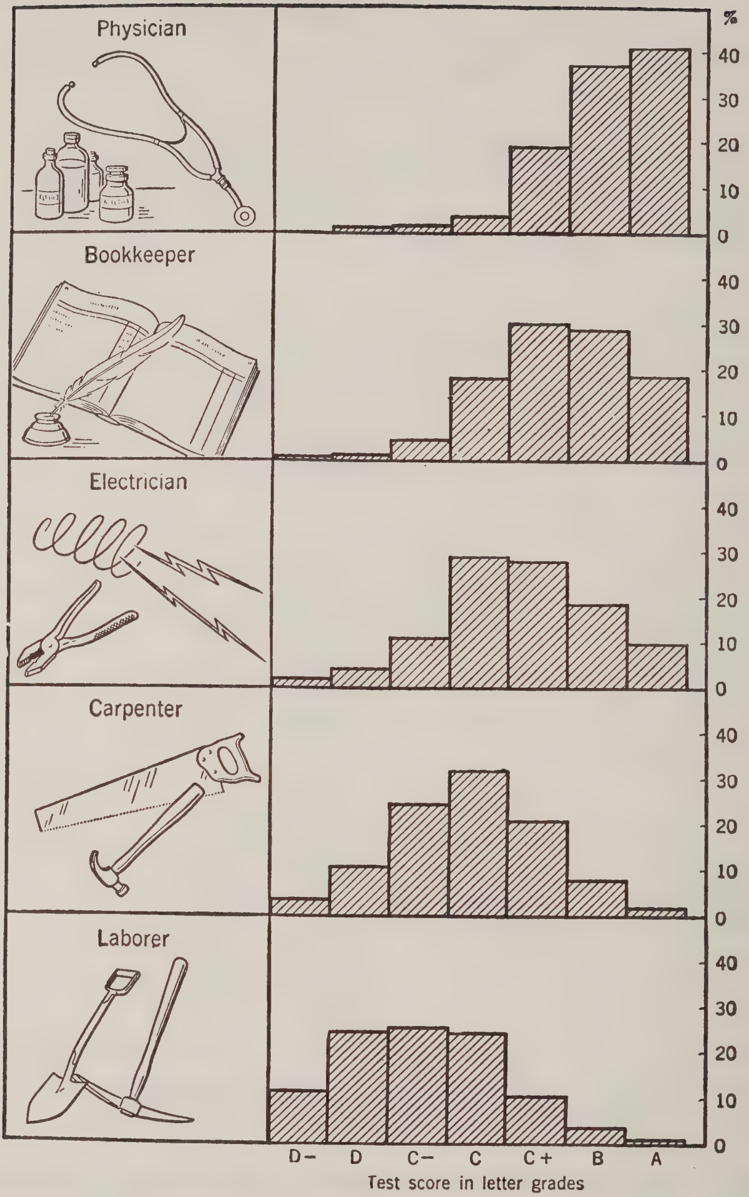


Figure 50. Distribution of Scores on the Army Alpha for Five Occupational Groups. (From Woodworth's *Psychology,* Fourth Edition. Henry Holt and Company, 1940, page 126.)

that pupils with an aptitude score of below 24 have a chance that exceeds 50 per cent of making a mark of *D* or failure. The scores on two or more tests[33] can be combined, with appropriate weights, if they are first expressed in standard scores or in some other common unit, and the probability of success for each of the combined scores can be represented in a similar table. An ordinary two-way table, or scatter diagram, which shows degrees of achievement on one axis associated with various test scores on the other axis, is also a useful table in guidance work.

TABLE 45

NUMBER AND PERCENTAGE OF FAILURES AND "*D*'S + FAILURES" AT EACH INTERVAL ON THE LEE TEST OF GEOMETRIC APTITUDE (AFTER LEE)

SCORE	NUMBER OF FAILURES	PER CENT OF FAILURES AT EACH LEVEL	NUMBER OF *D*'S + FAILURES	PER CENT OF *D*'S + FAILURES
1	*2*	*3*	*4*	*5*
68+				
64–67			2	11.1
60–63	1	3.0	4	12.1
56–59			3	8.3
52–55	2	4.9	4	9.8
48–51	2	6.5	4	12.9
44–47	2	4.9	6	14.6
40–43	3	7.3	6	14.6
36–39	5	10.6	14	29.8
32–35	2	7.1	11	39.3
28–31	8	32.0	11	44.0
24–27	5	14.7	17	50.0
20–23	7	36.8	13	68.4
16–19	9	50.0	15	83.3
12–15	8	44.4	13	72.2
8–11	2	66.7	3	100.0
4–7			1	100.0

In the final analysis, however, the success of the interview depends to a large degree upon the personality of the interviewer. For this reason counselors should be chosen with great care. An effort should be made to enlist persons who have had successful experience as classroom teachers, who are well liked by their pupils and genuinely interested in them, and who are intellectually and emotionally mature.[34] If counselors have this background and some special training, the counseling should be highly effective.

---

[33] For excellent discussions of the use and interpretation of tests in guidance, together with a description of available tests, see: Walter Van Dyke Bingham, *Aptitudes and Aptitude Testing,* 390 pages. New York: Harper & Brothers, 1937; and Donald G. Paterson, Gwendolen G. Schneidler, and Edmund G. Williamson, *Student Guidance Techniques,* 316 pages. New York: McGraw-Hill Book Company, Inc., 1938.

Placement. An important phase of vocational guidance is the placement of the individuals in some line of work. Many schools have some kind of bureau whose function is to contact prospective employers, to arrange for interviews, and to assist students in various ways in locating positions. This is the culmination of the guidance program. Effective placement can hardly be expected unless it is preceded by a careful program of preliminary guidance.

Follow-up. There are two major objectives of the follow-up phase of the guidance program.[35] Attention has been called to the fact that guidance is always in terms of probabilities, never in terms of certainties. Guidance attempts to aid the individual in making the best choice that is possible under the circumstances and in the light of the data available at the time. But both the individual and his environment will frequently change in ways not fully predictable. It is usually necessary, therefore, to regard any long-time planning as tentative in character and subject to such modifications as future developments may warrant. The guidance service should be continuous, not sporadic, and should extend beyond the period of formal schooling. It is always desirable, however, to wait until the program outlined has had a fair trial before altering it. One could hardly judge the superiority of the touch system over the two-finger system of typing at the end of a week's trial, or the ultimate success of a salesman by his record at the end of the first month. On the other hand, guidance counselors can render a most useful service to the individual by carefully watching his progress toward his educational and vocational goals, and by advising him when changes may be desirable. At the present time the follow-up aspect of the guidance program in most schools is very inadequate.

The follow-up is not only valuable to the pupil but is also a needed check upon the guidance program of the school. The value of the guidance service of the school should not be taken for granted, as is usually done. The only way to determine the value of guidance is to compare the future performance of those who have received and followed it with that of similar individuals in the same school who have not. A follow-up study of this type at the University of Minnesota, involving 987 students, indicates that 90 per cent of the students who carried out wholly or partly the recommendations of the Testing Bureau made satisfactory adjustment or progress toward adjustment, as compared with 22 per cent of the students who failed to follow the recommendations.[36] Another study by Webster [37] indicated that after two to five years, about 75 per cent of the predictions were correct, 13.8 per cent doubtful, and 11.2 per cent were incorrect.

---

[34] For a complete discussion of interview procedures, see: Walter Van Dyke Bingham and Bruce Victor Moore, *How to Interview,* 3d edition, 308 pages. New York: Harper & Brothers, 1941. Excellent briefer discussions are to be found in Ruth Strang, *Counseling Technics in College and Secondary School,* pages 52–82. New York: Harper & Brothers, 1937; *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I, op. cit.,* pages 126–136; and John G. Darley, *Testing and Counseling in The High-School Guidance Program,* pages 164–185. Chicago: Science Research Associates, 1943.

[35] For a good discussion of follow-up procedure, see Williamson and Darley, *op. cit.,* Chapter IX.

C. Vocational Guidance

Present program inadequate. In spite of the fact that vocational guidance was the first type of guidance to receive marked attention during the first decade of the twentieth century, there is good reason for thinking that the present program is quite inadequate. In 1937, Edgerton [38] conducted a follow-up study of 143 large and small communities located in 29 states. Although there was practically unanimous acceptance in theory that "one of the recognized major purposes of modern education is to aid young persons with their problems of self-inventory, self-discovery, and self-development . . . it was discovered that a majority of the 7,912 boys and girls studied will of necessity find their places in the worldly scheme of affairs largely as luck and accident happen to dictate." The specific limitations of the guidance service are indicated in the following statement: [39]

> The findings revealed, among other things, (1) that students are not giving enough thought to their educational preparation, occupational plans, recreational activities, and community contacts; (2) that students do not choose wisely of school offerings, life careers, health provisions, outside experiences, and work opportunities, when left to their own devices; (3) that students are most influenced in their educational, occupational, and personal choices by individuals who are not well qualified to advise them on such matters; (4) that the high school and the college do not function as they should in aiding students to decide upon their course of preparation, their choice of occupation, their program of recreation, their plan for employment, or their participation in other community situations.

---

[36] Donald G. Paterson, "The Genesis of Modern Guidance," *Educational Record,* 19: 44, January, 1938.

[37] Edward C. Webster, "A Follow-up on Vocational Guidance," *Journal of Applied Psychology,* 26: 285–295, June, 1942.

[38] A. H. Edgerton, "Guidance in Transition from School to Community Life," *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I,* pages 242–245. Bloomington, Illinois: Public School Publishing Company, 1938.

[39] *Ibid.,* page 243. Quoted by permission of the Society.

The same author also reported in 1938 a nationwide study of 2,630 separate occupational classifications and the corresponding training provisions made by the public schools. He found that even for the large number of pupils of low-level ability the program of studies was predominantly of the college-preparatory type. He also found that vocational programs in high schools and colleges were attempting to prepare pupils for occupational patterns of a type from five to twenty years earlier rather than for the essential present-day requirements. From these results Edgerton concluded that the following school and community opportunities for occupational adjustment are needed for aiding individuals:[40]

(1) To acquire from reliable sources up-to-date information about occupational trends, possibilities, and requirements,
(2) To check aptitude and personal quality ratings with actual specifications for corresponding positions in the locality,
(3) To secure unbiased counsel on tentative learning-earning plans from competent workers and employers,
(4) To test the wisdom of occupational choices through supervised experiences of self-discovery in the community,
(5) To obtain adequate preparation for occupational life through cooperative or apprenticeship training,
(6) To locate avenues of employment most conducive to personal growth, health, and happiness,
(7) To secure accurate information concerning supplementary preparation required for success and advancement,
(8) To pursue re-training activities under favorable conditions in case circumstances warrant such procedure.

The service that measurement can render to such a program will now be briefly pointed out.

Analysis of vocational opportunities and requirements. Until comparatively recently, vocational guidance made two errors in emphasis. First, it stressed information about occupations and neglected the study of the individual seeking guidance. It also stressed the opportunities afforded in terms of salary income and the like, and neglected to give specific information as to the abilities required for success and the probable demands for such services. Such statements as appear about the abilities required are "exceedingly vague and general or are based on arm-chair speculation."[41] Only quantitative statements regarding minimum amounts are helpful, for all occupations require some intelligence, industry, physical strength, and other traits. As stated by Toops,[42] the differentiating characteristics of an occupational group must be "unique, minimal in numbers, and as objective, quantitative and practical as possible."

---

[40] *Ibid.*, pages 244–245. Quoted by permission of the Society.

[41] Donald G. Paterson, Gwendolen G. Schneidler, and Edmund G. Williamson, *op. cit.*, page 277.

[42] Herbert A. Toops, "Some Concepts of Job Families and Their Importance in Placement," *Educational and Psychological Measurement*, 5: 195–216, Autumn, 1945.

In recent years, however, some useful beginnings have been made. Psychologists in England and in America have come to recognize the value of a quantitative picture of the particular pattern of abilities and interests which characterizes successful workers in different occupational groups and which can be applied to individuals considering possible entrance into these fields. In England, as a useful makeshift standard, ratings for each of 80 vocations on an elaborate occupational scale have been made and 23 minimum abilities and other qualities have been determined.[43] The United States Employment Service has prepared a valuable *Dictionary of Occupational Titles*[44] which describes more than 17,000 separate jobs and classifies each job with those to which it is closely related.

A helpful type of analysis is that which reveals the characteristics that differentiate the most successful from the least successful members of a given occupational group. For example, Anderson[45] found that 48 per cent of the "best" sales clerks in certain departments possessed "insight" and "well-integrated personalities," as against only 6 per cent of the "worst" sales clerks. Perhaps the most scientific attempt made so far in the measurement description of jobs in terms of the human abilities required is that of the Minnesota Stabilization Research Institute.[46] Gooch, however, made a study of 138 of the best available books and monographs written expressly on occupations, and came to the conclusion that the "tragic truth of the matter is that for 43,485,108 out of a total of 48,829,920 gainfully occupied persons in the United States, available occupational information is so inadequate as to be of little, if any, practical value to counselors, personnel workers, and students of occupations."[47]

The Committee on Social Trends has made a valuable analysis of the data in the United States Census of Occupations.[48] Figure 51, based on this analysis, and prepared by the Bureau of Agricultural Economics shows the shifts that took place between 1870 and

---

[43] Charles Allen Oakley and Angus Macrae, *Handbook of Vocational Guidance*, pages 130–137. London: University of London Press, 1937.

[44] The Superintendent of Documents, Washington, D. C.

[45] V. V. Anderson, *Psychiatry in Industry*, page 253. New York: Harper & Brothers, 1929.

[46] Donald G. Paterson and John G. Darley, *Men, Women, and Jobs, A Study of Human Engineering*, 145 pages. Minneapolis: University of Minnesota Press, 1936.

[47] Wilbur I. Gooch, "Occupational Information: Neglected Fields in the Available Literature," *Occupations*, 12: 34, March, 1934.

[48] Ralph G. Hurlin and Meredith B. Givens, "Shifting Occupational Patterns," in the one-volume edition of *Recent Social Trends in the United States*, pages 268–324. New York: McGraw-Hill Book Company, Inc.. 1933.

1940. It is evident that the increase in the percentage of persons employed in trade and transportation, clerical service, and professional service has been especially marked, while that in agriculture and allied occupations has declined. Another analysis[49] also indicates this trend.

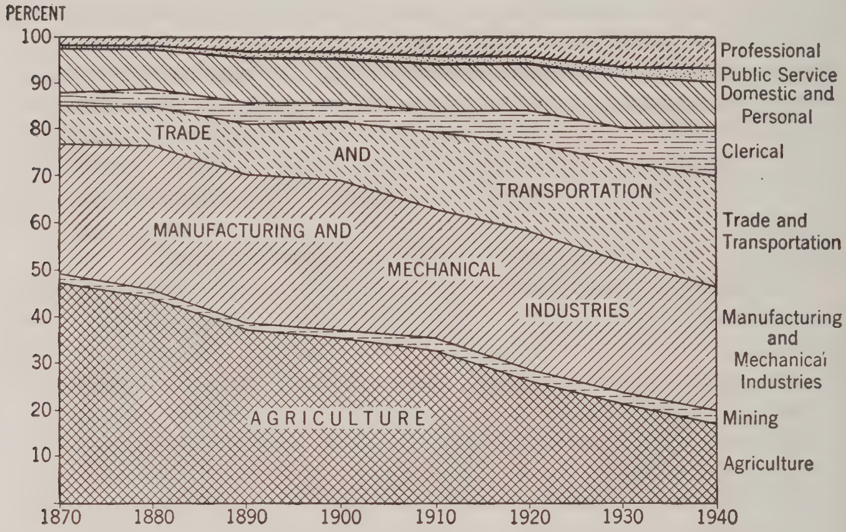


Figure 51. Shifts in Major Occupational Groups in the U. S. from 1870 to 1940.

Even more valuable is such an analysis for the locality of the school. Analyses of this sort have been made for several cities, including New York City, St. Paul, Minneapolis, and Duluth. When such an analysis is placed alongside that of the occupational preferences expressed by the pupils, the discrepancies are likely to be so marked as to provoke thoughtful discussion. Figure 52 shows the distribution of choices for ten occupations made by 1,000 boys in the 7*B* grade of a New York City school, compared with the number of men among 1,000 workers actually following these occupations in the city. That similar discrepancies also exist in the case of girls is indicated by a study made by Brown and Larson.[50]

---

[49] H. Dewey Anderson and Percy E. Davidson, *Occupational Trends in the United States,* pages 16–17. Stanford University: Stanford University Press, 1940.

[50] Clara M. Brown and Agnes A. Larson, *A Survey of the Working Experience and Future Plans of the Girls in the Secondary Schools of St. Paul in Relation to Various Educational and Economic Factors,* 44 pages. St. Paul: Minnesota Department of Education, 1938.

Clark [51] stresses the need of up-to-date information for each occupation regarding the wages, number employed, number needed immediately and at varying periods of time, present unemployment along with a ten-year average, and reliable information relating to probable technological changes to be expected.[52]

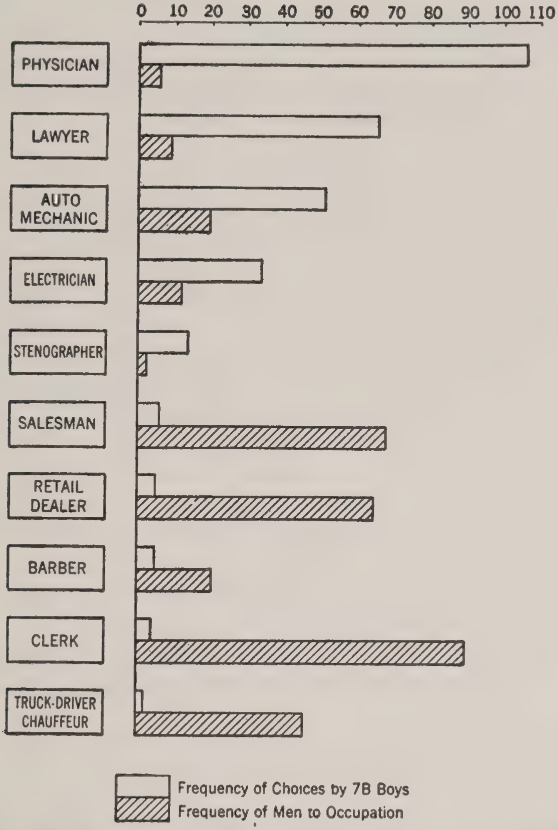


Figure 52. Distribution of Choices for Ten Occupations Made by 1,000 7B Boys in New York City Compared with the Number of Men Among 1,000 Workers Actually Following These Occupations in the City.

Analysis of the individual. Although it is true that individuals differ markedly in their ability for any particular occupation, authorities on vocational guidance now believe that it is possible for every person to be happy and reasonably successful in a number of vocations. Kitson, for example, says, "It is probable that 50 per cent of the people can succeed with a 50-percentile degree of success in 50 per cent of the occupations."[53] Along the same line Cunliffe says:[54]

---

[51] Harold F. Clark, "Planning Occupational Distribution," *Occupations,* 12: 18–26, February, 1934.

[52] The New National Occupational Information and Guidance Service in the United States Office of Education is an important step in this direction. See *School and Society,* 49: 50, January 14, 1939.

> Vocational guidance has been too much concerned with matching abilities possessed with those demanded. . . . Most people, so far as ability is concerned, are capable of doing a wide variety of things equally well, and with regard to abilities required, so far as we now have any means of knowing, wide ranges of occupations demand similar combinations of ability.

Because of the constant changes through which most occupations are going, it appears that adaptability, the capacity for learning and willingness to learn, is one of the most essential qualities. Edgerton notes "plenty of evidence to predict that in the job hunt of tomorrow the race will be to the socially well-adjusted and to the versatile."[55] Cunliffe thinks that, regardless of the nature of the work, the fundamental problem in most cases is one of adjusting the worker's personality to the occupational situation. He suggests that maladjustment, when it occurs, is usually due to one or more of the following factors:[56]

1. An inability to do the job—the small minority.
2. Social ineffectiveness, which is seen in
    (a) Failure to get on with superiors.
    (b) Failure to get on with inferiors.
    (c) Failure to work well with associates.
    (d) Inability to adjust oneself to the life-pattern of the occupation.
3. Misunderstanding of the true nature of the vocational world.
4. Lack of an intelligent philosophy of work.
5. A failure of the job rather than of the individual.

It is apparent, therefore, that intelligence, both abstract and social, is important, usually even more important than knowledge and skill, particularly at the time of employment. A high degree of social intelligence is always an asset, and the particular level of abstract intelligence demanded depends upon the occupation. Up to the present, however, the efforts of psychologists to measure the former have not been very successful. They have done a better job in measuring the latter. It is a well-known fact that it is possible for persons to have too much intelligence for a particular job, as well as too little. If the former is true, they are likely to be unhappy and dissatisfied; if the latter is true, they are likely to find that the time required to master the job is too great. Because this is the case, Fryer and Sparling assert that "with the exception of a try-out in the occupation itself, the general intelligence test is by far the best single predictive measure of success in an occupation." [57] Figure 49 indicates how these authors would use such data. It is often very important to be able to indicate to a person whether he would be one of the most intelligent or one of the least intelligent workers in the occupation being considered.

---

[53] Edwin A. Lee (Editor), *Objectives and Problems of Vocational Education,* page 260. New York: McGraw-Hill Book Company, Inc., 1938.

[54] Rex B. Cunliffe, *Trends in Vocational Guidance,* pages 14–15. New Brunswick, N. J.: School of Education, Rutgers University, 1935.

[55] *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I, op. cit.,* page 235. Quoted by permission of the Society.

[56] Rex B. Cunliffe, *op. cit.,* page 16.

Of course, other factors about the individual besides his level of general intelligence must be considered. In a number of fields, tests of specific intelligence, or aptitude, have been developed. Interest inventories, such as those by Strong,[58] Thurstone,[59] and Kuder,[60] when used with a person beyond the age of seventeen, apparently have their greatest value in indicating whether or not he will find a given occupation congenial and pleasant, granted he has the ability for it.[61] Frandsen [62] summarizes the experimental literature and concludes that "*something* about a person is measured quite reliably by interest inventories," but that such inventories "correlate negligibly with achievement, aptitudes, and possibly with curriculum satisfaction." A promising technique by which an individual may develop an interest in an occupation which comes within the range of his abilities has been presented by O'Rourke.[63] Bingham puts the situation as follows: [64]

> Self-knowledge is a gradual growth. To gain a clear understanding of one's aptitudes is an achievement of years rather than of hours. . . . As aids to self-understanding, scientifically constructed tests of aptitudes are not a substitute for insight and common sense. They may, however, serve to supplement or modify the considered judgment of a counselor who combines and weighs all the facts, from the personal history and the personal interview as well as from the test record.

---

[57] Douglas Fryer and E. J. Sparling, "Intelligence and Occupational Adjustment," *Occupations,* 12: 55, June, 1934.

[58] Published by Stanford University Press.

[59] Published by University of Chicago Press.

[60] Published by Science Research Associates.

[61] Walter Van Dyke Bingham, *op. cit.,* pages 70–82.

[62] Alden Frandsen, "Appraisal of Interests in Guidance," *Journal of Educational Research,* 39: 1–12, September, 1945.

[63] Harry D. Kitson, "Creating Vocational Interest," *Occupations,* 20: 567–571, May 1942.

[64] Walter Van Dyke Bingham, *op. cit.,* pages 12–13.

Counseling. By wise counseling the individual is assisted in thinking through his vocational problems until he arrives at what appear to be the best possible decisions under the circumstances. It is always a process of co-operative problem solving, with the pupil assuming as much responsibility as he is capable of. Logically the process will include the following steps,[65] although not necessarily in the order given:

1. The discovery and statement of the problem. Counselors soon find out that many students are not aware of their problems and the attendant implications;
2. The search for and statement of alternative courses of action;
3. The search for and use of all facts and knowledges that bear on the problems or the alternative solutions;
4. The consideration of all probable outcomes or end-results;
5. The decision—the selection of the course of action which in light of all the information secured will result in optimum adjustment and the construction of a program to that end.

Of course, the pupil may make mistakes. But human beings often learn much from their mistakes. "Experience, especially unpleasant experience, is often the best counselor, when all else fails." [66] Often a brief try-out experience will convince the pupil that he is on the wrong track. He is then ready for further counsel in the light of his new experience. Usually the data will warrant only tentative solutions. Before the pupil arrives at even a tentative decision, he should be led to ask himself such questions as the following: [67]

1. What level of general education is expected of people who enter this occupation? Have I the necessary schooling or can I acquire it?
2. In addition to the general schooling, how long a period of specialized education or training is ordinarily necessary? Where can I secure it, and what will it cost?
3. What level of intelligence has been found to characterize the people who enter upon and make progress in the occupation? Do my general mental abilities resemble those of persons in this field?
4. Are any special talents or aptitudes necessary? If so, are they a part of my endowment?
5. Specifically, what kinds of activity are most characteristic of this occupation? Do I like to do these kinds of things? Should I find the work and the surroundings congenial?
6. What are the average annual earnings of people in this occupation? At what rate should I start, and what income might I eventually expect? Are there exceptional rewards at the top?
7. Is employment relatively secure and steady, or intermittent, seasonal, hazardous?

---

[65] Rex B. Cunliffe, *op. cit.*, page 37

[66] E. G. Williamson and J. G. Darley, *op. cit.*, page 41.

[67] Walter Van Dyke Bingham, *op. cit.*, page 5.

8. What are the opportunities for advancement? Is this a blind alley, or does it open doors to other occupations?
9. What is the ratio of employment opportunities to the supply of competent applicants? How keen is the competition I should face?

Placement and follow-up. Apparently the most neglected aspects of the guidance program are placement and follow-up.[68] It has been estimated that only during the past decade as many as one in ten individuals leaving school and college have received any helpful follow-up supervision after leaving these institutions.[69] A recent book on guidance says: "It seems obvious that the next great step in vocational guidance must be more adequate placement facilities and the follow-up of all pupils who have attended public schools." [70] These services should include not only the purely vocational phases of after-school life but should assist in locating opportunities for desirable social-civic and recreational activities in the community.

The vocational-guidance program must be regarded as incomplete unless it succeeds in locating the individual in an occupation in which he has reasonable promise for success and happiness, and in which he makes maximum use of his talents. The prospective worker should be more concerned with the long-time opportunities afforded for developing a well-rounded, satisfied, and adjusted life, than with the initial wage. For this reason as much care should be given to studying the record and policies of the prospective employer as to the interest and ability of the applicant. It must be recognized that the final test of the guidance service and of the total educational program of the school is in terms of the quality of adjustment its pupils make to the economic, civic, and social life of the community.

The very complete follow-up program of Providence, Rhode Island, has been described by Allen.[71] In this city the counselors who served the pupils during the three-year stay in the senior high school attempt to check on them at intervals of one, three, and five years after they leave school. The counselors succeed in getting returns from practically all at the end of one year, from about 95 per cent at the end of the third year, and usually from about 85 to 90 per cent at the end of the fifth year. Such a check-up enables the school to extend its guidance service through the critical period of transition from the school to the occupational world, and at the same time enables the school to make any needed changes in its guidance program that appear desirable in the light of the follow-up studies.

---

[68] Rex B. Cunliffe, *op. cit.,* page 45.

[69] *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I, op. cit.,* page 246.

[70] Philip W. L. Cox and John Carr Duff, *Guidance by the Classroom Teacher,* page 129. New York: Prentice-Hall, Inc., 1938.

[71] Richard D. Allen, "Continuous Follow-up Survey in the Senior High School," *Vocational Guidance Magazine,* 10: 105–110, December, 1931.

The fullest possible co-operation between the school and the economic life outside is of the utmost value to all concerned. The employer would find that the reactions of the pupils, both on tests and in the interview, are more trustworthy when secured in the familiar environment of the school than when secured under the stress and strain of the employment office. Furthermore, he would be able to interpret the findings in comparison with other individuals in the same situation. The school would profit from this close contact, because it would better understand the demands of the employer and the conditions of work offered. The school should also cooperate fully with, and aid in every possible way, the nation-wide system of employment offices being developed under the United States Department of Labor.

D. Educational Guidance

Vocational versus educational guidance. In recent years the emphasis in guidance has shifted somewhat from vocational to educational guidance. This is certainly true of all grades below the senior high school.[72] A recognized leader in the industrial applications of psychology states this newer point of view in five basic principles, as follows: [73]

> 1. The basic problem in the school is one of *educational* guidance rather than *vocational* guidance. Nothing in our vocational experiments so far warrants the prediction of vocational success during the early years of education. A multitude of findings warrants our prediction of educational success. Vocational considerations in educational guidance, yes; but educational guidance primarily.
>
> 2. Educational guidance should be the *major* concern of our educational institutions, not an incidental or an *additional* (and easily subtracted) activity such as vocational guidance has been in the past.
>
> 3. The eighth or ninth grade pupil should know the educational requirements of typical occupations just as well as he knows the capital of Peru or the extraction of square root. The teaching of occupational information is beside the point and may be even harmful unless done in terms of the education and preparation which occupations require, *and the capacity* of the student to acquire such education.
>
> 4. Educational guidance, with its vocational considerations, should start from an analysis of the individual, not from a consideration of the economic system and its occupational eccentricities. The individual is the only constant variable in the complicated equation of society.

---

[72] William Martin Proctor, "Trends in Pupil Guidance," *California Journal of Secondary Education*, 10: 113–117, January, 1935.

[73] Henry C. Link, "Wheat and Chaff in Vocational Guidance," *Occupations*, 13: 11–12, October, 1934.

> 5. Educational guidance should be based on scientific tests of the individual's capacities and aptitudes. These instruments, as a means for predicting educational progress, are far better than their present use would indicate. . . .

The position that measurement occupies in relation to this guidance program will be considered under these topics:

1. Analysis of educational opportunities and requirements.
2. Analysis of the individual.
3. Counseling.
4. Placement and follow-up.

Analysis of educational opportunities and requirements. The problems in whose solution educational guidance attempts to be of help fall roughly into two groups. The first has to do with the choice of an educational program, and the second has to do with the choice of effective means of carrying it out. Such problems as the selection of a school or of a curriculum within a school illustrate the first type, while problems concerned with the use of the library and with the improvement of study procedures illustrate the second. Naturally the type of information required for an intelligent choice will depend upon the problem. While educational guidance should be a continuous process, its service is greatest at the important transitional points, such as from home to school, from the elementary school to the secondary school, from the secondary school to college, and from college to life outside. It is also valuable at the beginning of some new school subject, such as algebra or a foreign language.

Earlier in the chapter attention has been called to an error made in vocational guidance, which for a long time occupied itself with imparting information about vocations and neglected the analysis of the individual. A common error in educational guidance has been just the reverse of this; it has frequently been concerned with the analysis of the individual and has neglected to study the opportunities presented by the school situation. There can be little doubt that in schools and in educational programs there are individual differences that are quite as marked and as significant as the differences in the pupils themselves.

This fact is indicated by such reports as that of Kaulfers,[74] who surveyed fifty-one correlation studies used in the prognosis of foreign language achievement. He found, for example, that 132 correlations between foreign language achievement and such personal factors as intelligence, character traits, and chronological age varied from —.57 to .99, with a median of .356. Kaulfers concluded that the range of these coefficients shows "too great variability to warrant confidence," and that they "will probably always vary with differences in courses of study, methods of instruction, and nature of class personnel." Successful guidance, therefore, would appear to depend quite as much upon a knowledge of the traditions, organization, methods, standards, and ruling educational philosophy of the particular school or class (reflected to a considerable extent in the mortality rate), as upon the interests and abilities of the pupil. As a matter of fact, two leaders in college personnel work assert that, if they were permitted to choose the college, they could guarantee that "any high school graduate could emerge from it four years later with a baccalaureate degree." [75] Counselors who are familiar with the comparative distributions of intelligence in different colleges, such as appear in the *Educational Record* from year to year on the American Council Psychological Examination, would be in a favorable position to assist a prospective student in selecting an institution well suited to his ability. A recent study [76] shows that it is possible for students whose IQ is as low as 91–95 to obtain an A.B. degree from an institution with as high academic standards as Oberlin College.

---

[74] Walter Vincent Kaulfers, "Present Status of Prognosis in Foreign Language," *School Review,* 39: 585–596, October, 1931.

The necessary adjustment needed for success can often be secured by changing the school rather than the pupil. There is no convincing reason to think that the school program is fixed and that only the pupil is subject to change. Guidance workers, as well as other educators, appear often to forget the fact that the school program is made for the pupil, and not the pupil for the program. Good [77] has prepared a helpful volume primarily for the use of veterans' counselors.

Seashore has stated admirably what should be the aim of education and its relation to guidance, in these words: [78]

> The educational objective which underlies all scientific guidance is that *it is the function of the educator to keep each child busy at his highest natural level for successful achievement in the field for which he has reasonable aptitude and in which he will find a reasonable outlet for self-expression, in order that he may be happy, useful and good.* We have not yet reached more than a verbal acceptance of this undeniable principle either in music or general education; but it is our

---

[75] *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I, op. cit.,* page 68. See also *Phi Delta Kappan,* 26: 6, September, 1943.

[76] L. D. Hartson, "Influence of Level of Motivation on the Validity of Intelligence Tests," *Educational and Psychological Measurement,* 5: 273–283, Autumn, 1945.

[77] Carter V. Good (Editor), *A Guide to Colleges, Universities, and Professional Schools in the United States,* 697 pages. Washington: American Council on Education, 1945.

[78] Carl E. Seashore, "Educational Guidance in Music," *School and Society,* 45: 386, March 20, 1937.

inevitable goal. The main thing that is blocking its acceptance is the lack of an acceptable and thorough-going guidance program as a part of the educational system.

Analysis of the individual. A distinguished educational historian suggests that it is fairly easy to find out how much a candidate for college *has learned,* somewhat difficult to find out how much more he *can learn,* and almost impossible to prophesy how much more he *will learn.*[79] While this statement is doubtless relatively true, the first two parts appear to err slightly in the direction of optimism, and the last part in the opposite direction. After a comprehensive survey of the extensive research in predicting college achievement, Strang [80] came to this conclusion:

> There are three factors which seem to determine scholastic achievement to the greatest extent—intelligence, previous educative experience, and purpose. The various standard tests of intelligence are probably the best available measure of potential ability, and the achievement and placement tests, of specific preparation. High-school marks, in addition to testing these two qualities, may be the best criterion of sustained purpose over a period of years.

The later studies reported make it unnecessary to alter this conclusion, except perhaps to emphasize the fact that these data are much more useful when accessible on the cumulative record card for a number of years than when secured at the end of the high-school period or at entrance to college. The same three factors operate in much the same way in predicting achievement in the high school, although the correlations there are somewhat higher. The predictions made in advance are closer for the freshman year than for the entire high-school or college period, and for average marks than for marks in the individual subjects.

Abundant experimental evidence supports the conclusion that at all educational levels the best single prediction that can be made of an individual's record in a subject is his past record in that subject or in closely related subjects.[81] "The best prediction of an individual's future success in a given area is the level of his achievement in that area up to the present." [82] Almost without exception the closest correlation with the second year's record in any high-school or college subject is the first year's record in that subject. Generally the second semester's record can be predicted from the first semester's record in the subject better than from any test of general intelligence or of specific aptitude. Also, the records of very superior and of failing students can be predicted much better than can those of students of average performance. A combination of two factors gives predictions which are both higher and more stable than those based on a single factor. Adding still more factors raises the correlations only slightly. Odell,[83] in a very comprehensive study of freshmen in about 100 small colleges, found that the high-school average alone gave a correlation of .55 with the freshman average and the Otis IQ alone gave .38, and that the two combined raised the correlation to .58.

---

[79] I. L. Kandel, *Examinations and Their Substitutes in the United States,* page 57. New York: Carnegie Foundation for the Advancement of Teaching, 1936.

[80] Ruth Strang, *Personal Development and Guidance in College and Secondary School,* pages 72–133; 188–237. New York: Harper & Brothers, 1934.

[81] For a good survey of the literature, see: Saul B. Sells, "Measurement and Prediction of Special Abilities," *Review of Educational Research,* 14: 38–54, February, 1944.

[82] Giles M. Ruch and David Segel, *op. cit.,* page 40.

At times special factors may operate to make one's past record a poor index of his probable future achievement. There is evidence that this is often the case with veterans who have returned to school after three or more years in the Army or Navy. Even the performance of such individuals on psychological tests administered immediately upon their return to school may be untrustworthy. The special circumstances must be taken into account.[84]

In two important studies Hartson [85] compares the relative merits of ratings by high-school principals, teachers, and friends with high-school scholarship records and the Ohio State University Psychological Examination as indices of academic achievement. Table 46 summarizes the principal findings of the second study, which in essential respects confirms the first. It will be noted that in predicting college achievement the intelligence test, which is one of the best available, slightly exceeds the high-school record and considerably exceeds any simple average of the eight ratings. Of the ratings, only the first three appear to have much predictive value; those on emotional stability and appearance actually operate to reduce the correlations. When the first three ratings are combined by the multiple-correlation procedure, the prediction is approximately as good as that of high-school scholarship. When ratings on these three traits are combined with the high-school scholarship, the multiple is approximately as high as the simple correlation of the intelligence test with first-semester scholarship at Oberlin. It will also be noted that ratings by the principal on most traits are usually somewhat higher than those by the high-school teachers

---

[83] Charles W. Odell, *Predicting the Scholastic Success of College Freshmen,* Bureau of Educational Research Bulletin, No. 37, 54 pages. Urbana: University of Illinois, 1927.

[84] Loren S. Hadley, "To What Extent Will Colleges Adjust to the Needs of Veterans?," *School and Society,* 63: 323–325, May 4, 1946; see also *Educational Research Bulletin,* 24: 87–92, 112, April 18, 1946.

[85] See *School and Society,* 36: 413–416, September 24, 1932; and 46: 155–160, July 31, 1937.

(given in parentheses). The ratings by friends, not included here, are very much inferior. Note that the predictions are uniformly higher for men than for women at Oberlin.

TABLE 46

CORRELATION OF RATINGS BY HIGH-SCHOOL PRINCIPAL AND TEACHERS, OHIO STATE UNIVERSITY PSYCHOLOGICAL EXAMINATION, AND HIGH-SCHOOL SCHOLARSHIP, WITH FIRST-SEMESTER SCHOLARSHIP AT OBERLIN AND WITH HIGH-SCHOOL SCHOLARSHIP (ADAPTED FROM HARTSON)

VARIABLE	COLLEGE SCHOLARSHIP		HIGH-SCHOOL SCHOLARSHIP	
	Men	*Women*	*Men*	*Women*
Ratings by Principal and Teachers:[a]				
1. Intelligence	.47(.47)	.37(.34)	.58(.55)	.51(.49)
2. Industry	.33(.35)	.28(.22)	.47(.43)	.37(.34)
3. Attitude	.39(.36)	.38(.33)	.53(.48)	.45(.47)
4. Work habits	.36(.40)	.27(.23)	.48(.49)	.42(.40)
5. Reliability	.32(.28)	.19(.19)	.36(.35)	.23(.19)
6. Leadership	.27(.19)	.18(.07)	.36(.28)	.31(.23)
7. Emotional stability	.29(.28)	.17(.12)	.31(.32)	.19(.18)
8. Appearance	.23(.16)	.05(–.02)	.20(.10)	.14(.11)
1–3, total	.50(.51)	.45(.41)	.77(.72)	.64(.65)
1–4, total	.49(.51)	.44(.40)	.76(.73)	.64(.65)
1–5, total	.49(.50)	.42(.39)	.63(.62)	.56(.55)
1–8, total	.49(.47)	.38(.30)	.60(.57)	.52(.49)
Psychological examination	.61	.57	.40	.38
High-school scholarship	.58	.51	—	—

[a] Ratings by teachers are in parentheses. In 85 per cent of the cases the rating is the average of two teachers.

Quaid[86] has called attention to a limitation of many studies of prediction: they are too general. He suggested that such studies ought to show what happens to the boy as distinguished from the girl, and to the bright student as compared with the dull, in the particular school in which the information is to be used in guidance. He found that in Phillips University the average high-school marks tended to predict college freshman marks better for boys than for girls, to be superior to general intelligence tests for abler boys, and to be inferior to them for abler girls and less able boys. He observed a tendency for girls to exceed expectation more often than boys, and for boys to fall behind expectation more often than girls. These findings indicated strongly the superiority of specific predictions over general predictions. Rundquist[87] has found a closer agreement between school marks and intelligence test scores for girls than for boys in the junior high school, but not in the elementary school.

---

[86] T. D. D. Quaid, "A Study in the Prediction of College Freshman Marks," *Journal of Experimental Education,* 6: 350–375, March, 1938.

The comparative value of standardized objective tests and teachers' marks in prediction appears to depend somewhat upon the local situation. Read,[88] for example, found a correlation of .63 between high-school averages and first-semester freshman marks at the Municipal University of Wichita, where more than two thirds of the students enter from a single school system, as compared with correlations of .44, .42, .42, and .41, respectively, for the Iowa High School Contest Examination, the Ohio State University Psychological Examination, the Iowa Silent Reading Test, and the Purdue Placement Test in English. As a rule, colleges drawing students from many systems find high-school marks of about the same value as, or only slightly superior to, objective tests, except tests of personality, which are usually much lower.

One practical difficulty with the high-school average and with intelligence tests administered in the senior year is that the information comes too late to be of maximum value for guidance. Byrns and Henmon[89] found, however, that practically as good predictions of achievement at Wisconsin University could be made from IQ's obtained in the tenth grade and from the average tenth-grade mark. They also found that IQ's from the National Intelligence Test obtained when pupils were in grades four to eight of the Madison schools correlated with first-semester marks at the University about as well as did the psychological test percentiles. But here again local conditions seem important. Adams[90] found essentially negative results in predicting high-school and college records in Texas from IQ's on the National Intelligence Test and scores on the Stanford Achievement Test administered in grades four to six. Rosenfeld and Nemzek[91] also found no value in IQ's from the Detroit First Grade Intelligence Test for predicting marks at Wayne University. It is apparently much safer to rely upon test scores and school marks for the period just preceding the one for which prediction is sought, although tentative predictions are possible much earlier. Measurement as well as guidance must be a continuous process and not restricted to any one period.

---

[87] Edward A. Rundquist, "Sex, Intelligence, and School Marks," *School and Society,* 53: 452–456, April 5, 1941.

[88] See *School and Society,* 48: 187–188, August 6, 1938.

[89] Ruth Byrns and V. A. C. Henmon, "Long-Range Prediction of College Achievement," *School and Society,* 41: 877–880, June 29, 1935.

[90] See *Journal of Educational Psychology,* 29: 56–65, January, 1938.

[91] See *School and Society,* 47: 127–128, January 22, 1938.

While it is generally true in high school as in college that the best prediction of achievement during the first year can be made by combining an individual's previous school record reflected in teachers' marks with scores on tests of general intelligence, other data are important, particularly for certain specific purposes. For example, the pupil's chronological age is very important in predicting his persistence in school. Ross [92] found that the age at which pupils completed the eighth grade yielded correlations with the number of semesters spent in high school that varied from —.50 to —.61 for four different classes. Thorndike [93] found that the grade reached at the ages of 14, 15, or 16 years, when combined with the age to which his family plans to keep a pupil in school, correlates .90 or better with the grade actually reached. Maller [94] studied the records of 5,783 seniors in six high schools located in four states. He found that the correlation between chronological age and scholarship varied from —.35 to —.48, as compared with .28 to .41 between scholarship and scores on the Terman and Otis group tests of intelligence. For predictive purposes, negative coefficients are as good as positive coefficients of the same magnitudes.

For a few high-school subjects, aptitude tests seem to be the best single basis for predicting achievement. These subjects usually represent fields the content of which is least like that of the elementary school. Examples of such subjects are art, music, industrial arts, algebra, geometry, and foreign languages.

The information which is of greatest value in educational guidance is suggested in the following list, in approximate order of merit:

1. The complete record of the pupil's school experience to date, especially his recent achievement, and his chronological age.
2. The complete record of all standard tests.
3. Estimates by the principal and teachers as to his ability, attitudes, industry, and work habits.
4. Statements by the pupil regarding his educational and vocational goals, interests, likes, and dislikes.
5. The pupil's health record and present health status.
6. The pupil's family history, especially the educational and economic status of the family.

Counseling. It should be kept in mind that the correlations in the foregoing section represent conditions where little or no guidance was available. Presumably, intelligent guidance would have directed many of these pupils into fields where their academic performance would have been better. Inadequate guidance is reflected in a heavy mortality rate both in high school and in college. Even more serious is evidence that the most promising pupils often drop out of school prematurely while the least promising ones continue. For example, Thorndike [95] studied the records of 785 eighth-grade boys in New York City and found that of the 40 especially able boys 5 left school before the age of 15, but not one of the 40 especially weak ones did. He also found that for every boy in the top 40 who stayed in school beyond the age of 18, nearly 10 of those below average ability did so. Strang [96] summarizes studies showing similar results in state-wide surveys in Kansas, Ohio, and Wisconsin. One of the most important social values of guidance is the securing of a better distribution of education.

---

[92] Clay Campbell Ross, *The Relation between Grade School Record and High School Achievement,* pages 63–66. New York: Bureau of Publications, Teachers College, Columbia University, 1925.

[93] Edward L. Thorndike and associates, *Prediction of Vocational Success,* page 113. New York: The Commonwealth Fund, 1934.

[94] J. B. Maller, "Age versus Intelligence as Basis for Prediction of Success in High School," *Teachers College Record,* 33: 402–415, February, 1932.

The United States Office of Education [97] has recently compiled data which indicate that under present conditions about 20 to 25 per cent of secondary-school pupils should eventually find themselves in college preparatory courses, about 10 to 25 per cent in courses aiming at the skilled trades, and at least 50 per cent in the more general type of secondary education. The plan suggested is that at the beginning of the secondary-school period all pupils enter upon a period of broad general education and continue in this program until they have demonstrated that they are qualified through aptitudes, interests, abilities, and desires to branch off into special fields leading to the professions or to skilled occupations. It is only through a greatly improved guidance program that this distribution can be effected.

Special care should be exercised in interpreting low scores, however. Retests are often desirable. Even when low scores are confirmed by later tests, other considerations must be taken into account. It must be remembered that the correlation that can at present be obtained between the optimum combination of predictive factors and subsequent achievement is far from perfect. A pupil may be discouraged but should rarely be denied a trial at the program desired. An important role of the counselor is to inform the pupil regarding the odds involved in the choice.

Both the pupil and his parents have a right to know the objective data upon which the recommendation is based. But this does not mean that pupils' exact scores on intelligence and aptitude tests must be given, although at times this knowledge may be desirable. As a rule, however, all that is necessary is to state that the pupil is relatively low or relatively high, in the upper half of his class or in the lowest fourth of his class, and the like. In other words, report the data "descriptively," as is done at the University High School, Oakland, California, rather than "numerically."[98] Some evidence has been reported in an earlier chapter which indicates that the relatively low pupils profit greatly from such information. In the elementary school a valuable basis for guidance is *internal;* that is, in relation to the pupil's own strong and weak points and his own past record. But in the high school, the standard, at least in part, must be *external;* that is, the individual must learn to appraise himself in relation to others as he begins to think seriously about a vocation in life. Bingham states the situation well in the following words:[99]

---

[95] Edward L. Thorndike, "The Distribution of Education," *School Review,* 40: 335–345, May, 1932.

[96] Ruth Strang, *Personal Development and Guidance in College and Secondary School, op. cit.,* pages 85–86.

[97] Testimony before the Temporary National Economics Committee, Washington, D. C., April 25, 1940. See also *School Life,* 28: 6, July, 1946.

> But testing stops short of performing its most vital service unless its outcome is revealing, not only to the educational institution but also to the student himself, giving him more definite assurance as to what he really wants and needs—the kinds of problems he should now bite into, the topics for which his appetite is ready, the subjects of study toward which he should reach, the sort of college course or other advanced training he would like to pursue, yes, eventually, the calling and the way of life in which he anticipates finding his fullest self-realization. A battery of tests has missed the bull's-eye of its target unless the students learn from it something true and significant about themselves.

Placement and follow-up. The placement of the pupil must always be regarded as tentative, and subject to whatever changes later developments may appear to warrant. In the words of Ben D. Wood:[100]

> The highest rule of measurement in education is the prophecy of long-term provisional goals for individual pupils, and the progressive modification of these goals in accordance with cumulative evidence of growth and of needs, intellectual, personal and social.

At the present time, however, there can be little doubt that the follow-up program is the weakest link in the guidance chain, as well as one of the most important links. Seashore speaks of the "generous testing and the very meager and inadequate follow-up work which is a common curse today."[101] Eurich and Wrenn regard this attitude as comparable to the "architect's withdrawing after the blueprints are made or to the doctor's losing interest once a prescription is written." [102] The responsibility of the counselor ends only when the individual has made a satisfactory adjustment to the new situation.

---

[98] Marion Brown and Vibella Martin, "Techniques Used in Guidance at University High School," *University High School Journal,* 14: 23–44, June, 1935.

[99] *Educational Record,* 20: 138, Supplement No. 12, January, 1939.

[100] *Test Service Bulletin,* No. 35, page 5. Yonkers: World Book Company, 1935.

[101] *School and Society,* 45: 393, March 20, 1937.

E. Personal Guidance

Student problems. Some of the most troublesome problems that students face can be classified as neither vocational nor educational, although often closely related to both. They are better described as personal, for indeed many of them are of a distinctly intimate or personal character. Table 47 gives a list of 784 problems, classified into six categories, presented by 196 University of Minnesota students.[103] Many of these problems will require the service of specialists. Problems involving health and physical disabilities will usually be referred to the department of health for correction. Many problems involving serious social and emotional maladjustments will best be handled by specialists in mental hygiene or by counselors with psychiatric training in addition to a thorough grounding in psychology. The ordinary guidance workers can be of great help, however, in locating persons who have such problems and in assisting them in finding competent professional advice. Merely "talking over" the less serious personal problems with a sympathetic counselor, who is himself emotionally well-adjusted, will be of considerable help in objectifying and intellectualizing the problem. Frequently the counselor can help to work out a better relation with the student's parents and the general social environment.

Value of personality measurement. The need for valid measuring instruments in the diagnosis of personality weaknesses is fully as great as in any other type of diagnosis. But unfortunately few such instruments exist. Although Traxler [104] estimates that about 500 tests and inventories of personalities have been published, he concludes that they are "for the most part . . . still definitely experimental." After a comprehensive survey of existing tests and scales, Lee arrives at these conclusions: [105]

---

[102] *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I, op. cit.,* page 78. Quoted by permission of the Society.

[103] E. G. Williamson and J. G. Darley, *op. cit.,* page 257.

[104] Arthur E. Traxler, *op. cit.,* pages 98–99. Brief but useful descriptions of the most promising instruments are in Chapter VI.

[105] J. Murray Lee, *A Guide to Measurement in Secondary Schools,* page 170 New York: D. Appleton-Century Company, 1936.

Table 47

TYPES OF PROBLEMS PRESENTED BY 196 UNIVERSITY STUDENTS (after WILLIAMSON and DARLEY)

Types of Problems	Frequency
1. *Financial:* Need or desire for part-time work, scholarship or loans; inadequate finances	45
2. *Vocational:* Poor aptitude for chosen vocation; choice of an occupation; dearth of interest in any vocation; dearth of interest in chosen vocation; information about occupations; vocational choice without adequate self-analysis; inadequate information in regard to professional choice	300
3. *Educational:* Poor aptitude for college work; selection of courses in line with occupational choice; inferiority in academic skills such as reading, study habits, English usage, etc.; understanding grading standards; high general aptitude and poor scholastic achievements; understanding responsibilities in college; high aptitude hampered by standard curricula; outside work interfering with studies; university entrance without proper requirements	227
4. *Social, Personal, and Emotional:* Too much social life or too many social activities; inadequate participation in extracurricular activities; selecting student activities in line with interests; social personality traits which may hinder professional success; need for encouragement and self-confidence; social timidity; emotional disturbances; family domination in vocational choice; conflict with family or friends; parental anxiety for a wise vocational choice; fear of intellectual inadequacy; idealization of a profession; overevaluation of a college degree; vocational indecision because of possible marriage; no self-expression in the home	136
5. *Family:* Sibling conflict; split-family situation; over-dependence on part of student; difficulty in transition to independence; serious and overt family conflict over educational and vocational plans, finances, religion, difference in standards, personality adjustment	39
6. *Health and Physical Disabilities:* Serious physical disabilities; easily fatigued; inability to do justice to work because of intermittent illness; physical habits, diet and sleep, etc.	37
Total	784

Rating scales will probably prove of the greatest all-around value to the school. . . .

Moral knowledge and honesty tests should usually only be used as a necessary part of some larger investigation planned by a trained research worker. . . .

Those situations where facilities are such that definite use can be made of results. . . . Teachers should be given results only if and when it is necessary to enlist their aid.

Attitude scales may be more widely used. . . . Remedial work should be handled with tact and by an indirect rather than a direct approach.

Feder and Mallett[106] report the study of 21 cases of personality maladjustment at the University of Iowa with three well-known scales, the Thurstone Personality Schedule, the Woodworth-House Mental Hygiene Inventory, and the Bell Adjustment Inventory. Upon the basis of the findings they conclude that "paper-and-pencil personality questionnaires of the types herein studied have doubtful validity for the discovery and diagnosis of personality maladjustments."

A recent comprehensive summary[107] concludes that "group-administered paper and pencil personality questionnaires are of dubvious value in distinguishing between groups of adjusted and maladjusted individuals, and that they are of much less value in the diagnosis of individual adjustment of personality traits." However, fifteen studies in which the tests were administered individually reported more encouraging results. When used as individual tests rather than as group tests the results were positive in ten studies, questionably positive in three, and negative in only two studies.

In 1939 Ruch and Segel summarized the situation with respect to rating scales as follows:[108]

> The experimental literature on the value of rating scales shows that the tide of confidence in such measures ebbs and flows. Twenty years ago the World War gave a great impetus to the rating idea. This movement was followed by a decade of critical study in which evidence accumulated that raised grave questions about the validity and reliability of the method. Today there is a swing back toward confidence in ratings, partly at least because rating scales have been greatly improved.

Even if the existing tests and scales are not so valid as could be desired, they are still of some value in supplying information preliminary to the interview. The behavior of the individual during the interview and in the test situation is often very revealing. Allen has developed a useful technique for group guidance, which very effectively combines the self-measurement of groups with individual interviews on a voluntary basis.[109] A slightly abridged form of Project 58, given below, illustrates this technique, which is applicable to the measurement of personality as well as to the measurement of achievement and special abilities. Boyer[110] has described the self-appraisal program employed in the junior high schools of Philadelphia in which each pupil records his own scores on a chart and plots his profile on nineteen measures of aptitude, basic skills, interests, and adjustments.

---

[106] Daniel D. Feder and Donald R. Mallett, "Validity of Certain Measures of Personality Adjustment," *Journal of the American Association of Collegiate Registrars,* 13: 5–15, October, 1937.

[107] Albert Ellis, "The Validity of Personality Questionnaires," *Psychological Bulletin,* 43: 385–440, September, 1946.

[108] Giles M. Ruch and David Segel, *op. cit.,* page 30.

[109] Richard D. Allen, *Self-Measurement Projects in Group Guidance,* 274 pages. New York: Inor Publishing Company, 1934.

[110] Philip A. Boyer, "Group Testing in the Philadelphia Public Schools," *Education* 66: 416–423, March, 1946.

SOCIAL ADJUSTMENT (GRADES 7 TO 10)[111]

PROBLEM: How do I compare with others in my ability to get along with other people?

I. *Preparation of the Counselor*

1. Objectives: To help counselors to discover and assist pupils who have antisocial tendencies; to help the pupils themselves to discover unhealthy symptoms in their attitudes towards personal and social problems; and to produce an attitude of guidance readiness in regard to problems involved in developing social attitudes.
2. References: The Manual for the Personality Index Test by Loofbourow and Keys.[112]
3. Materials: Copies of the Personality Index Test for each pupil, the Manual of Directions, and keys for scoring.

II. *Suggestions for Motivation*

1. Explanation: Assume that the test is a regular examination, similar to those which the pupils have previously taken.

III. *Administration of the Test*

1. Directions: Follow carefully the directions in the Manual and on the test papers.
2. Scoring: Follow carefully the directions for scoring provided in the keys. Note that a high score is obtained in proportion to the number of *wrong* answers.
3. Statistical Treatment of Results: When pupils have totaled their scores, ballots should be distributed on which pupils should write their scores but not their names. Fold, collect, and arrange the ballots in order of score; distribute scores on the board in convenient intervals; find the score of the median pupil, and distribute the scores in deciles which, in turn, may be converted into letter ratings from A to E in a ratio of 1–2–4–2–1. Let each pupil translate his score into a letter rating.
4. Diagnostic Study: Summarize each part of the test separately with ballots as for the total examination. Notice Test 1 is really a test of honesty in schoolwork; Test 2, of social attitudes; Test 3, of conduct; and Test 4, an inventory of personality and attitudes. The combined scores of the battery constitutes the *personal index.*

IV. *Issues and Implications*

1. Discussion: Most children, as they pass through school and grow more mature, learn how to get along with other people. They learn how to observe the rules at school, at home, and on the playground; they learn to respect the property and rights of other people; and they learn how

---

[111] *Ibid.,* pages 223–226.

[112] Prepared by G. C. Loofbourow and Noel Keys and published by Educational Test Bureau, 1933.

to conduct themselves in such a way that will win the approval of their fellows. Some students, however, seem to have the tendency to get into trouble frequently. The cause of such difficulties may be discovered to some extent by the test which we have just taken. Notice the range of scores in this test. The highest tenth of the scores are A ratings; the lowest tenth of the scores are E ratings; the next two tenths at either end are B and D ratings, respectively, while the middle four tenths are C ratings. What should an A or B rating indicate? It would be interesting to know, although we shall not ask, how many pupils with A or B scores frequently get into difficulty at home, at school, or on the playground. What should a D or E score indicate?

Compare ratings on the separate tests. What would it mean if your rating was very high on Test 1? On Test 2? On Test 3? On Test 4? Are there any who rate high in some of these tests and low in others, or is your general level high or low in all? What would such distribution of scores indicate? What can a person do about it if his score is very high in the test as a whole, or in any part of it? Construct individual profiles. Is it a matter for individual effort towards a solution, or are there some ways in which the counselor or parents may help? What measures would you suggest for a person with an A rating in order to improve his attitudes towards social problems? What part does self-interest or selfishness play? A desire to get something for nothing or without effort?

2. Invitation to an Individual Interview: Those who would like to discuss the results of this test with the counselor are invited to do so. Many times there are suggestions which the counselor can offer, especially in regard to certain sections of the test, that will help pupils in overcoming handicaps. Any pupil who wishes an interview is invited to let the counselor know about it.

V. *Possible By-Products*

1. A Continuous Survey: The early discovery of pupils with character, conduct, or personality difficulties is an important part of the junior-high-school program. If pupils can be led to appreciate their own difficulties and to seek the assistance of the counselor, much may be done to correct such handicaps.
2. Effects of the Project on Attitudes and Standards: If pupils voluntarily come to the counselor because they appreciate their personality and conduct difficulties, much can be done to help them. Even if a pupil himself does not seek an interview with the counselor, the presence of such difficulties in the class will have been detected, and, in many cases, the counselor will soon be able to put names and facts together in such a way that difficulties can be identified and individuals sought out for interviews. The general effect of the project should be to cause the individuals who need warning to feel that the problem applies to them personally. The discussion is more important than the validity of the test, since it is for group-guidance purposes rather than clinical diagnosis.

SELECTED READINGS FOR FURTHER STUDY

Bell, Howard M., *Matching Youth and Jobs.* Washington, D. C.: American Youth Commission, 1940. 274 pages.

Bingham, Walter Van Dyke, *Aptitudes and Aptitude Testing.* New York: Harper & Brothers, 1937. 390 pages.

Chisholm, Leslie L., *Guiding Youth in the Secondary School.* New York: American Book Company, 1945. 433 pages.

Crawford, Albert Beecher, and Burnham, Paul Sylvester, *Forecasting College Achievement: A Survey of Aptitude Tests for Higher Education.* New Haven: Yale University Press, 1946. Three Volumes.

Darley, John G., *Testing and Counseling in the High-School Guidance Program.* Chicago: Science Research Associates, 1943. 222 pages.

Erickson, Clifford E., and Happ, Marion Crosley, *Guidance Practices at Work.* New York: McGraw-Hill Book Company, Inc., 1946. 325 pages.

Fryer, Douglas, *The Measurement of Interests in Relation to Human Adjustment.* New York: Henry Holt & Company, 1931.

Germane, Charles E., and Edith G., *Personnel Work in High School.* New York: Silver Burdett Company, 1941. 599 pages.

Kefauver, Grayson N., and others, *Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I.* Bloomington, Illinois: Public School Publishing Company, 1938. 313 pages.

Lyons, G. J., Martin, H. C., and Lynch, J. T., *The Strategy of Job Finding.* New York: Prentice-Hall, Inc., 1940, 408 pages.

Myers, George E., *Principles and Techniques of Vocational Guidance.* New York: McGraw-Hill Book Company, Inc., 1941. 377 pages.

Paterson, Donald G., Schneidler, Gwendolen G., and Williamson, Edmund G., *Student Guidance Techniques.* New York: McGraw-Hill Book Company, Inc., 1938. 316 pages.

Rappaport, David, *Diagnostic Psychological Testing.* Chicago: Year Book Publishers, Inc., 1946. Volume I and Volume II.

Rogers, Carl R., *Counseling and Psychotherapy.* Boston: Houghton Mifflin Company, 1942. 450 pages.

Smith, Charles M., and Roos, Mary M., *A Guide to Guidance.* New York: Prentice-Hall, Inc., 1941. 440 pages.

Strang, Ruth, *The Role of the Teacher in Personnel Work.* New York: Bureau of Publications, Teachers College, Columbia University, 1935. 417 pages.

————, *Personal Development and Guidance in College and Secondary School.* New York: Harper & Brothers, 1934. 341 pages.

Symonds, Percival M., *Psychological Diagnosis in Social Adjustment.* New York: American Book Company, 1934. 362 pages.

Traxler, Arthur E., *Techniques of Guidance.* New York: Harper & Brothers, 1945, 394 pages.

————, *The Use of Tests and Rating Devices in the Appraisal of Personality.* New York: Educational Records Bureau, 1938. 80 pages.

Williamson, E. G., *How To Counsel Students.* New York: McGraw-Hill Book Company, Inc., 1939. 562 pages.

————, Darley, J. G., and Paterson, Donald G., *Student Personnel Work.* New York: McGraw-Hill Book Company, Inc., 1937. 316 pages.

————, and Hahn, M. C., *Introduction to High School Counseling.* New York: McGraw-Hill Book Company, Inc., 1940. 314 pages.

CHAPTER XVII

Evaluation

A. The Problem of Evaluation

Measurement and evaluation. As used in education, *evaluation* is a far more inclusive concept than *measurement.* Two aspects of evaluation may be distinguished: (1) data relating to some important aspect of the school, such as its organization program, or results; and (2) a set of values or standards against which these data are interpreted and appraised. Furthermore, the evaluator's educational philosophy and sense of values will determine what objectives of the school program he considers to be important, as well as what data he will look for, or regard as relevant in the situation. It is apparent that while measurement may be highly mechanical and at times a routine, evaluation can never be; at every stage evaluation requires the exercise of mature judgment.

Measurement implies the use of some tool or instrument, such as a test or scale, and provides a quantitative description of observed phenomena. This is always desirable, but it should never exclude relevant data of a subjective and qualitative character, or the consideration of outcomes not immediately observable. Some writers[1] have criticized existing measurement in education for the reason that it furnished inadequate data for evaluation. At best, measurement merely provides data needed for evaluation; it is not evaluation per se.

The *Sixteenth Yearbook*[2] of the Department of Elementary School Principals of the National Education Association presents a good discussion of evaluation on the elementary level. The ten chapters of this report are given below:

I. The Fundamentals of School Appraisal
II. Appraising the School Organization
III. Appraising Administrative and Supervisory Procedures
IV. Evaluating the Curriculum
V. Appraising Methods of Learning and Teaching

---

[1] *Cf.* Verner M. Sims, "Educational Measurement and Evaluation," *Journal of Educational Research,* 38: 18–33, September, 1944.

[2] "Appraising the Elementary-School Program," *The National Elementary Principal,* 16: 227–655, July, 1937.

VI. Evaluating Socializing Experiences
VII. Measuring the Progress of Pupils
VIII. Estimating the Efficiency of Teachers
IX. Judging School Equipment
X. A Review of the Technics of Appraisal

Note that the term "measuring" occurs in only one chapter heading. The other terms, "appraising," "evaluating," "estimating," and "judging," all similar in meaning, imply the use of techniques that go beyond testing and examining.

TABLE 48

THE MAIN METHODS OF EVALUATION USED BY THE COOPERATIVE STUDY OF SECONDARY SCHOOL STANDARDS, WITH THE WEIGHT ASSIGNED TO EACH

METHOD			PER CENT
1. Evaluative Criteria			40
A. Educational Program		20	
Curriculum	2.8		
Pupil activity	2.8		
Library	2.8		
Guidance	2.8		
Instruction	6.0		
Outcomes	2.8		
B. Organization and Plant		20	
Staff	10.0		
Administration	6.0		
Plant	4.0		
2. General Judgments by Visiting Committees			20
3. Growth as Measured by Standard Tests			20
4. Success of Pupils			10
A. In College	10 to 1		
B. Noncollege	0 to 9		
5. Judgment by Pupils			6
6. Judgment by Parents			4
Total			100

The Eight-Year Study of the Progressive Education Association,[3] which includes both elementary and secondary education, and the Three-Year Study of the Commission on Teacher Education [4] on the college level are splendid illustrations of an enlarged conception of evaluation. The committee has sought to devise suitable instruments of measurement for outcomes—such as interests, attitudes, creativeness, and various aspects of thinking—less tangible than those measured by ordinary tests and examinations. It has also utilized other types of data, such as anecdotal records, family histories, records of the pupil's activities, and the like.

---

[3] Eugene R. Smith, Ralph W. Tyler, and Evaluation Staff, *Appraising and Recording Student Progress,* 550 pages. New York: Harper & Brothers, 1942.

[4] Maurice E. Troyer and C. Robert Pace, *Evaluation in Teacher Education,* 368 pages. Washington: American Council on Education, 1944.

Possibly no more ambitious example of this enlarged conception of evaluation is available than the Cooperative Study of Secondary School Standards.[5] Table 48 shows the six main methods the study employed, of which only one represents the use of measurement in the ordinary sense.

The importance of evaluation. Without some form of evaluation everything about education becomes a matter of blindly hoping that all is well. In the critical period shortly before the Civil War, Abraham Lincoln began an important address with this statement: "If we could first know where we are and whither we are tending, we could better judge what to do and how to do it." It is no less true in education than in government that we must first "know where we are," and especially "whither we are tending," before we are in a position to judge intelligently regarding "what to do and how to do it." In the final analysis, it is the function of all attempts at evaluation to afford a basis of rational action. Apparently educators have always recognized this, even if often somewhat vaguely. For example, a college president said not long ago: "Self-criticism and self-appraisal (now a 'self survey' or an 'evaluation') are as old as education."[6]

The emphasis today is more and more upon the importance of *self*-evaluation. This holds true for all levels of education from the activity of the pupils in an elementary class passing judgment upon the success of a unit of instruction planned and executed by themselves to the formal Report of the Harvard Committee.[7]

More than a third of a century ago Thorndike pointed out that "the actual changes wrought in boys and girls by this or that form of education are being measured, old and new methods are being tested by experiment in the same spirit of zeal and care for the truth that animates the man of science, and the educational customs which have been accepted unthinkingly by 'use and wont' are being required to justify themselves to reason."[8] Although it is probably true that more progress has been made in that direction since the statement above was written than in all the centuries preceding, improvements in evaluation procedures have hardly kept up with those in curricula and teaching methods.[9]

---

[5] For a full account of this study, see *Evaluation of Secondary Schools, General Report,* 526 pages. Washington, D. C.: Cooperative Study of Secondary School Standards, 1939. For a briefer statement, see *How to Evaluate a Secondary School* (1940 Edition), 139 pages. Washington, D. C.: Cooperative Study of Secondary School Standards, 1939.

[6] Henry M. Wriston, "A Critical Appraisal of Experiments in General Education," *Thirty-Eighth Yearbook of the National Society for the Study of Education, Part II,* page 303. Bloomington, Illinois: Public School Publishing Company, 1939. Quoted by permission of the Society.

[7] *General Education in a Free Society,* 267 pages. Cambridge: Harvard University Press, 1945.

The difficulty of evaluation. The problem of evaluating education is immensely complicated. Many approaches toward a solution have been made, and none has been entirely satisfactory. For many years the various regional associations attempted to evaluate the secondary schools and colleges of America indirectly by their *possessions* rather than directly by their *products.* Such measures as the size and qualifications of the staff and the number of books in the library were at best indications of *educational opportunity;* and even to a less extent were such things as the number or type of buildings in the school plant and the amount of financial support available. The limitations of such a procedure have been characterized as follows: The standards used were mechanical, rather than vital; rigid, rather than flexible; deadening, rather than stimulating; traditional, rather than progressive; academic, rather than liberal; broadly comprehensive and subjective, rather than scientific.[10] Intensive and extensive study of the problem by competent committees within the past decade has increasingly revealed its complexity. The Cooperative Study of Secondary School Standards, for example, extended over a period of six years and cost about a quarter of a million dollars. It employed the six major methods of evaluation given in Table 48, and developed three scales, whose composition is given in Table 49. It will be noted that the complete scale, Alpha, includes 110 different "thermometers," all relating to the nine evaluative criteria of the first method listed in Table 49. In 1942 Eurich, Pace, and Ziegfeld [11] surveyed the literature of the field and came to this conclusion: "No simple and inexpensive technique has as yet been devised nor is one likely to be devised that will provide an evaluation of an entire educational program."

Evaluating teaching efficiency. A single illustration will show the complexity of the problem of evaluation. How can one best judge the worth of any particular classroom teacher? This is manifestly an important question. To a large extent the selection, growth, and promotion of teachers depend upon the answer. In general, the methods used are of three types. In the first place, tests and rating scales have been devised for measuring the *personality* of the teacher. As a rule, these have proved disappointing. The difficulty with this approach has been clearly pointed out by McCall: "No one has demonstrated just what causal relationship, if any, exists between possession of these various attributes and desirable changes in pupils." [12]

---

[8] Edward L. Thorndike, *Education, A First Book,* pages 7–8. New York: The Macmillan Company, 1912.

[9] *Cf.* Pedro D. Orata, "Evaluating Evaluation," *Journal of Educational Research,* 33: 641–661, May, 1940.

[10] *Evaluation of Secondary Schools, General Report, op. cit.,* pages 53–55.

[11] Alvin C. Eurich, C. Robert Pace, and Edwin Ziegfeld, "Evaluative Studies," *Review of Educational Research,* 12: 521–533, December, 1942.

TABLE 49

COMPOSITION OF 1940 EDITION OF THE ALPHA, BETA, AND GAMMA SCALES FOR EVALUATING SECONDARY SCHOOLS

AREA	NUMBER OF THERMOMETERS		
	Alpha	*Beta*	*Gamma*
Curriculum and course of study	19	8	3
Pupil activity program	13	5	2
Library service	11	7	3
Guidance service	7	3	2
Instruction	6	3	2
Outcomes	18	7	2
School staff	18	9	6
School plant	11	4	3
School administration	7	4	2
Total	110	50	25

A second method attempts to measure the worth of the teacher by her *performance,* usually her activity before the class. For this purpose various score cards and rating scales have been devised. In fact, the typical rating scale attempts to secure various measures of the teacher's performance, together with measures of certain traits of personality deemed important in teaching. But except as instruments of self-analysis by the teachers themselves, the practical value of rating scales is slight. For example, when the gains on the Stanford Achievement Test from November to May by pupils in four Wisconsin schools were used as a criterion, the correlations with 17 of the best-known measures of teaching ability available, although somewhat inconsistent, with few exceptions were so low that they could reasonably be supposed to have arisen from a population in which the true relationships were zero.[13]

---

[12] William A. McCall, *Measurement,* page 403. New York: The Macmillan Company, 1939.

A third method has been the attempt to judge the worth of the teacher by her *product,* the performance of her pupils. This is certainly the most direct, and is often asserted to be the only valid, approach. The most obvious way to achieve this result is to measure the improvement made by the pupils during a period of instruction under the teacher. But the problem is far more complicated than it at first appears. Even when allowances are made for differences in the intelligence and initial achievement of the pupils, the greater problem remains of determining how much of the growth is due to natural maturity and how much to the total educational environment in school and out of school, and the still greater problem of knowing how much of this improvement is due to the influence of any particular teacher. Most competent observers today would agree with Traxler[14] that "the use of test results for rating teachers is seldom advisable."

A recent summary by Barr[15] notes encouraging progress but emphasizes that the road ahead is long and difficult:

> The influence of any particular teacher is deeply enmeshed in a host of other school, pupil, and community factors. While very definite progress has been made in this area, it is not easy to isloate the effects of particular teachers in particular situations. There is reason to be optimistic about the use of more precise instruments of measurement in the management of the teaching personnel, but for the time being, discretion is the best part of valor.

The Cooperative Study. Doubtless, the most ambitious attempt at evaluation by means of standard tests has been the Cooperative Study of Secondary School Standards, involving 198 schools and a total of over 300,000 tests.[16] In spite of unusual care to avoid the difficulties summarized in Table 50, the Cooperative Study concluded that since the results showed that better methods of evaluation were available for accreditation, the use of standard tests should be restricted to diagnostic and guidance purposes by the local school. The Cooperative Study also attempted to judge the product of the school by follow-up studies of the subsequent careers and success, both academic and nonacademic, of former pupils, and concluded that this method was mainly of value for local school use. A periodic canvass of the opinion of pupils about the instruction and other

---

[13] Helen M. Walker (Editor), *The Measurement of Teaching Efficiency,* pages 73–141. New York: The Macmillan Company, 1935.

[14] Arthur E. Traxler, *Techniques of Guidance,* page 186. New York: Harper & Brothers, 1945.

[15] A. S. Barr, "The Use of Measurement in the Management of Teacher Personnel," *Education,* 66: 431–435, March, 1946.

[16] *Evaluation of Secondary Schools, General Report, op. cit.,* Chapter VIII.

Table 50

USE OF TESTS IN EVALUATING SCHOOLS[17]

Criticism	Comment
1. The ability of students varies widely in different schools.	Each pupil's achievement can be compared with others of the *same level of scholastic ability,* and not with the usual national norms.
2. A general testing program must be uniform and inflexible and does not recognize individual differences in schools.	Difficult to meet this objection. However, there is a large body of instructional material common to all institutions. This common core can be used as a basis of comparison by testing, leaving the differentiating phases to other types of evaluation.
3. A general testing program tends to crystallize curricula and to reduce instruction to mere coaching for examinations.	Danger is real, if tests are used for this purpose at regular intervals announced in advance. Does not hold for occasional testing to be used with other criteria. No objection to use of tests by schools for self-examination.
4. Available tests do not adequately measure important outcomes of instruction.	Undoubtedly true. But less true than formerly. Moreover, many important outcomes are measured by the better tests, and others are difficult to evaluate by any techniques so far developed.
5. The achievement of students at any given time is the result of *all previous schooling,* not merely that of the present school.	Can be met by using equivalent tests *before* and *after* a period of instruction, and by judging the worth of the school in terms of the *changes effected* between tests.
6. The average achievement of an institution as a whole does not properly take into account differences *within* the institution in curricula and departments.	A valid objection to *any single criterion* used for evaluating a school, and no more true of tests than of any other measure. Apparently the only answer is to present a picture of the *profile* of the institution showing the strong and weak points.
7. Measuring the status of the pupils at any given time tends to reflect the quality of the school at some time in the past, whereas what is wanted is a picture of the school as it is now functioning.	Can be met at least in part by measuring the *growth* produced during an instructional period in the school being evaluated.
8. Some measurable outcomes may be due to out-of-school contacts and so cannot properly be attributed to the influence of the school.	The objection probably holds mainly to the field of the social studies. In a sense, the degree to which this occurs is a measure of the success of the school, which should attempt to utilize and co-ordinate *all instructional agencies* that make up the educational environment, *out* of school as well as *in* school.
9. It is difficult to obtain standard conditions for administering a test to a variety of schools scattered over a wide area.	The difficulty can be reduced to a minimum by employing a small staff of carefully trained examiners who follow a simple program fully worked out in advance.

---

[17] Adapted largely from *Evaluation of Secondary Schools, General Report,* Chapter VIII.

aspects of the school which they are attending is also a valuable means of self-analysis and guidance; for, although the customer may not always be right, *what he thinks* about the institution is important, even when he is mistaken. In the Cooperative Study, pupil judgment also proved to be about as useful for evaluating schools as the elaborate testing program.

B. General Principles of Evaluation

For elementary schools. The Research Division of the National Education Association formulated the following statements of guiding principles for evaluating the programs of elementary schools; most of these appear equally applicable to the other levels of education: [18]

1. Adequate appraisal of the school includes more than the usual program of achievement testing.
2. School appraisal should be diagnostic; that is, it should reveal the specific points of strength and of weakness in the school program.
3. Every aspect of the school program should be appraised, regardless of its relative difficulty.
4. Principals and teachers should play important parts in the appraisal of their own schools. Their responsibility for planning and initiating appraisal measures will vary according to the plan of organization and administration in the school system as a whole.
5. Within reasonable limits and under proper safeguards, pupils also may contribute to school appraisal.
6. Evaluation of the school program should be carried on continuously. Pertinent information should be collected thruout each year and summarized at least once a year.
7. Methods of appraisal should be selected on the basis of their reliability, practicability, and appropriateness in the particular situation to be appraised. A combination of several methods is often better than one alone.
8. Before undertaking an appraisal, principals and teachers should find out how competent workers elsewhere have evaluated similar elements of the school program.
9. Careful subjective judgments formed in the light of valid criteria are better than conclusions based on objective data from a poorly planned or carelessly executed experiment.
10. Every appraisal should be made with reference to specified criteria of some kind. Such criteria should themselves be carefully evaluated before they are used.
11. Of the several types of criteria which may be used, those concerned with pupil development should receive first consideration.
12. There should be close agreement between the accepted objectives of a school and the instruments which the school uses to measure its attainment of these objectives.
13. When it is impracticable to determine the merits of local school practises directly, these practises should be appraised with reference to the findings of available research studies and expert opinion outside the school.

---

[18] *The National Elementary Principal, op. cit.*, 16: 237–238, July, 1937.

14. Statistical technics for determining the reliability of experimental results should be used only with a thoro understanding of their purpose and significance.

15. The results of appraisal should be used to improve the school program. It is essential that classroom teachers, as well as principals, be fully informed of these results.

16. Parents and pupils also should be given accurate information concerning the strong and weak elements of the school program, so that they may help to improve it.

For secondary schools. The Cooperative Study of Secondary School Standards has prepared the following eighteen principles,[19] which provide a comprehensive philosophy not only for evaluating secondary schools, but also for evaluating other levels of education:

1. American secondary schools, much as they may differ in details, are essentially alike in their underlying purposes and organization.
2. In a democracy the fundamental doctrine of individual differences is as valid for schools as for individuals. Schools, as well as pupils, differ from each other markedly.
3. A school can be studied satisfactorily and judged fairly only in terms of its own philosophy of education, its individually expressed purposes and objectives, the nature of the pupils with whom it has to deal, the needs of the community which it serves, and the nature of the American democracy of which it is a part. All American schools, however they may differ in type, have this in common: they are instrumentalities for transmitting our American heritage and our American democratic ideals. Provided this aim can be clearly kept in view in every case, each school is free to determine its own educational policies in promoting the ideals of American civilization.
4. A school should be judged in terms of the extent to which it meets satisfactorily the needs of all pupils who should come to it, not alone of those who continue their formal education in institutions of higher learning.
5. Methods of accreditation and interpretation of evaluation should recognize the differences in background, development, and existing conditions in different states and regions. No attempt should be made to develop uniform standards for the nation or to have them administered from a single national office.
6. It is more significant to measure what the school does than what it is or what it has. The educational process and product are more important to evaluate than the machinery and equipment.
7. A school should be judged as a whole, not merely as the sum of its separate parts.
8. The number of factors evaluated in the modern secondary school should be sufficiently large and varied to give valid evidence of the worth of the school in each of its main areas.
9. Accrediting criteria and procedures should be brief enough in extent, sufficiently varied in form, and convenient enough in arrangement to be practicable for use in all secondary schools.
10. Methods of evaluation and accreditation, as far as possible, should be based upon scientific studies and objective evidence, rather than upon untested assumptions and unsupported opinions.

---

[19] *Evaluation of Secondary Schools, General Report, op. cit.,* pages 57–61; also *How to Evaluate a Secondary School* (1940 Edition), *op. cit.,* pages 17–21.

11. The considered judgment of competent educators is an essential factor in the evaluation of the quality and character of the work of a school.

12. A valid method of evaluation and accreditation, based tentatively upon existing research studies and expert judgment, should be fully tested by extensive experimental try-out in a large group of typical, representative secondary schools throughout the country. The results of this experimentation should be carefully analyzed and evaluated.

13. While it is desirable in many respects that definite standards or levels of achievement should be developed, it is recognized that in most of the important aspects of a school's work the best available basis for the development of useful standards will probably be comparison with the practices in other comparable schools.

14. A good school is a growing school. It should be judged by its progress between two different dates as well as by its status at a single date.

15. Any useful, stimulating, and valid method of accreditation should be flexible with the passage of time; that is, it should be capable of reasonable modification as new bases of evaluation and different levels of achievement are suggested or developed from the use of existing ones.

16. If criteria for evaluation are sufficiently flexible, extensive, and thorough, it is not essential that they be applied annually.

17. The bases and methods of evaluation should be such as to require active participation in the process on the part of the entire professional and non-professional staffs of the school.

18. An important function of a national, regional, or state agency should be stimulation toward continuous growth and improvement, not merely inspection and admission to membership.

For higher institutions. The Committee on Revision of Standards created by the Commission on Higher Institutions of the North Central Association of Colleges and Secondary Schools spent five years and $135,000 in making a study reported in a series of seven monographs.[20] Section VI, entitled "Institutional Purposes and Clientele," and regarded as "the very heart of the new accrediting policy," is as follows: [21]

Recognition will be given to the fact that the purposes of higher education are varied and that a particular institution may devote itself to a limited group of objectives and ignore others, except that no institution will be accredited that does not offer minimal facilities for general education, or require the completion of general education for admission.

Every institution that applies for accreditment will offer a definition of its purposes that will include the following items:

1. A statement of its objectives, if any, in general education.
2. A statement of the occupational objectives, if any, for which it offers training.
3. A statement of its objectives in individual development of students, including health and physical competence.

This statement of purposes must be accompanied by a statement of the institution's clientele showing the geographical area, the governmental unit, or the religious groups from which it draws students and from which financial support is derived.

---

[20] *The Evaluation of Higher Institutions,* published by University of Chicago Press.

[21] George F. Zook and M. E. Haggerty, *Principles of Accrediting Higher Institutions,* pages 150–151. Chicago: University of Chicago Press, 1934.

The facilities and activities of an institution will be judged in terms of the purposes it seeks to serve.

C. Evaluating Various Aspects of the School

The philosophy of the school. There is rather remarkable agreement among the foregoing principles for evaluating the three levels of education, but nowhere is the agreement more notable than upon the point that an institution must be appraised in terms of its own philosophy and objectives. The Cooperative Study, for example, recommends that "a secondary school be studied expressly in terms of its own philosophy of education, its individually stated purposes and objectives, the nature of the pupils with whom it has to deal, the needs of the community which it serves, and the nature of the American democracy of which it is a part."[22] It recognizes four distinct phases in the satisfactory evaluation of a secondary school:[23]

1. Statement by the school of its philosophy of secondary education and of its objectives.
2. Checking and validation of the statements of philosophy and objectives against the needs of the pupil population and community which the school serves.
3. Revision or modification of the statements of philosophy and objectives, if necessary, in light of step number 2 above.
4. Evaluation of all aspects of the school in terms of these revised statements of philosophy and objectives. This phase involves the use of the rest of the *Evaluative Criteria.*

Figure 53 illustrates one of the procedures suggested for formulating the school's philosophy. Step 2 above indicates briefly the procedure to be followed in evaluating this philosophy, which is largely a matter of checking it for clearness, for internal consistency, and for appropriateness to the community to be served. Regarding the pupils and the community, the basic data which are required for the external evaluation are as follows:[24]

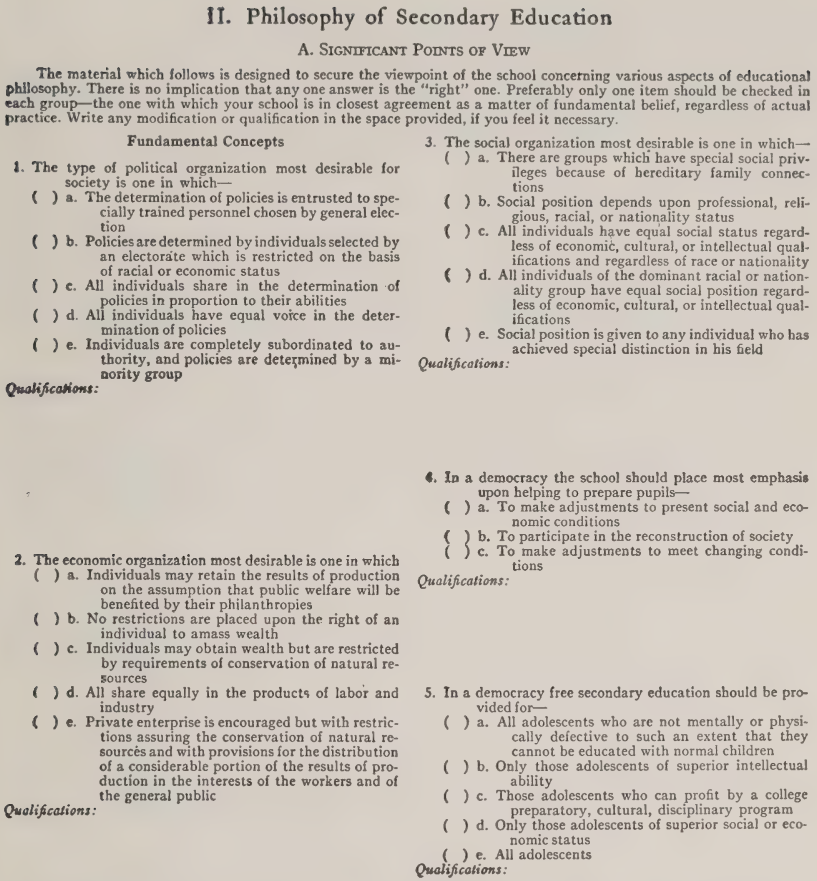
I. Basic data regarding pupils
   A. Graduates and enrollment by grades and by sex
   B. Number of years seniors have been in the school
   C. Distribution of withdrawals according to cause
   D. Age-grade distribution of pupils
   E. Distribution of I.Q.'s by grades

---

[22] *How to Evaluate a Secondary School* (1940 Edition), *op. cit.,* page 65.

[23] *Evaluative Criteria,* (1940 Edition), page 6. Washington, D. C.: Cooperative Study of Secondary School Standards, 1939.

[24] *Ibid.,* pages 17–28.



II. Philosophy of Secondary Education

A. Significant Points of View

The material which follows is designed to secure the viewpoint of the school concerning various aspects of educational philosophy. There is no implication that any one answer is the "right" one. Preferably only one item should be checked in each group—the one with which your school is in closest agreement as a matter of fundamental belief, regardless of actual practice. Write any modification or qualification in the space provided, if you feel it necessary.

Fundamental Concepts

1. The type of political organization most desirable for society is one in which—
   - ( ) a. The determination of policies is entrusted to specially trained personnel chosen by general election
   - ( ) b. Policies are determined by individuals selected by an electorate which is restricted on the basis of racial or economic status
   - ( ) c. All individuals share in the determination of policies in proportion to their abilities
   - ( ) d. All individuals have equal voice in the determination of policies
   - ( ) e. Individuals are completely subordinated to authority, and policies are determined by a minority group

Qualifications:

2. The economic organization most desirable is one in which
   - ( ) a. Individuals may retain the results of production on the assumption that public welfare will be benefited by their philanthropies
   - ( ) b. No restrictions are placed upon the right of an individual to amass wealth
   - ( ) c. Individuals may obtain wealth but are restricted by requirements of conservation of natural resources
   - ( ) d. All share equally in the products of labor and industry
   - ( ) e. Private enterprise is encouraged but with restrictions assuring the conservation of natural resources and with provisions for the distribution of a considerable portion of the results of production in the interests of the workers and of the general public

Qualifications:

3. The social organization most desirable is one in which—
   - ( ) a. There are groups which have special social privileges because of hereditary family connections
   - ( ) b. Social position depends upon professional, religious, racial, or nationality status
   - ( ) c. All individuals have equal social status regardless of economic, cultural, or intellectual qualifications and regardless of race or nationality
   - ( ) d. All individuals of the dominant racial or nationality group have equal social position regardless of economic, cultural, or intellectual qualifications
   - ( ) e. Social position is given to any individual who has achieved special distinction in his field

Qualifications:

4. In a democracy the school should place most emphasis upon helping to prepare pupils—
   - ( ) a. To make adjustments to present social and economic conditions
   - ( ) b. To participate in the reconstruction of society
   - ( ) c. To make adjustments to meet changing conditions

Qualifications:

5. In a democracy free secondary education should be provided for—
   - ( ) a. All adolescents who are not mentally or physically defective to such an extent that they cannot be educated with normal children
   - ( ) b. Only those adolescents of superior intellectual ability
   - ( ) c. Those adolescents who can profit by a college preparatory, cultural, disciplinary program
   - ( ) d. Only those adolescents of superior social or economic status
   - ( ) e. All adolescents

Qualifications:

Figure 53. A Suggested Technique for Evaluating the Philosophy of a Secondary School. (From *Evaluative Criteria*, 1940 Edition, Cooperative Study of Secondary School Standards, Washington, page 8.)

- F. Educational intentions of seniors by sex
- G. Occupational intentions of seniors by sex

II. Basic data regarding the community
- A. Population data for the school community
- B. Occupational status of adults
- C. Occupational status of youth of secondary school age
- D. Educational status of adults
- E. Financial resources of the school district
- F. Agencies affecting education
- G. Additional socio-economic information (seven items)

The educational program. A satisfactory statement of the school's philosophy having been formulated, a basis is now available for evaluating the educational program and organization of the school. The general point of view and procedure of the Cooperative Study is given in the "Instructions" reproduced in Figure 54.

The following list of educational "temperatures" indicates the comprehensive character of the evaluation of the educational program:

1. Curriculum and Course of Study.

   General principles; curriculum development; amount of offerings; English; ancient languages; modern languages; mathematics; sciences; social studies; music; arts and crafts; industrial arts; homemaking; agriculture; business education; health and physical education for vocational shop; general evaluation.

2. Pupil Activity Program.

   Nature and organization; school government; home rooms; school assembly; school publications; music activities; dramatics and speech; social life; physical activities of boys; physical activities of girls; school clubs; finances; general evaluation.

3. Library Service.

   Library staff; organization and administration; book collection, number of titles; book collection, recency; book collection, general adequacy; periodicals; supplementary materials; selection of materials; teachers and the library; use by pupils; general evaluation.

4. Guidance Service.

   Nature and organization; guidance staff; information about pupils; guidance procedures; phases of guidance; results; general evaluation.

5. Instruction.

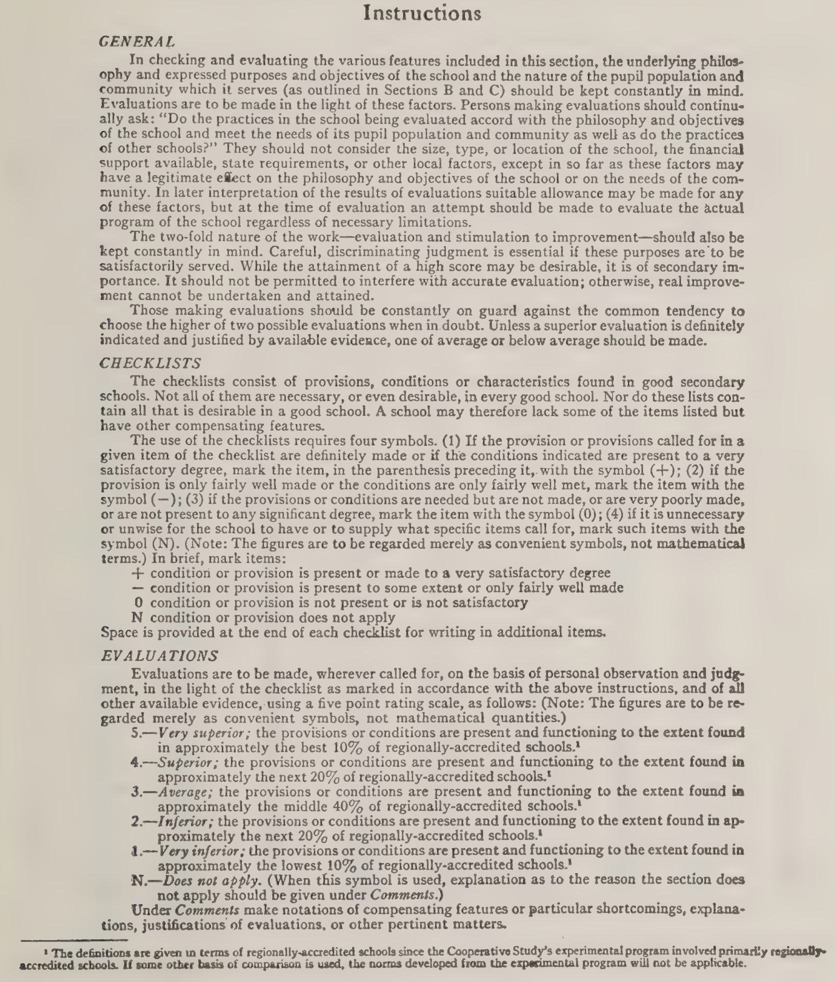
   Classroom activities; use of community; textbooks; methods of appraisal; special committee judgment; general evaluation.

6. Outcomes.

   Evaluation procedures; attainment in the principal subject matter fields; attitudes and appreciations.

Figure 55 illustrates the graphical summary of these "temperatures" for the median school. It will be noted that this school, which happens to be a large public school, is rated "average" (between the 30th and 70th percentiles) in four of the six areas of the educational program. The school is only at the 11th percentile, or "inferior," in the curriculum, however, and at the 80th percentile, or "superior," in library. The graphical device employed makes it possible to see at a glance the strong and weak points of the program.

Figure 56 shows the checklist and evaluations proposed for the content of the offerings in the principal subject-matter fields. Of these, only the two with the double stars at the bottom of the columns are included in the short Gamma Scale of 25 thermometers. These two are also included in the Beta Scale, together with the three additional fields indicated by single stars. The others appear only in the complete Alpha Scale of 110 thermometers.



Instructions

GENERAL

In checking and evaluating the various features included in this section, the underlying philosophy and expressed purposes and objectives of the school and the nature of the pupil population and community which it serves (as outlined in Sections B and C) should be kept constantly in mind. Evaluations are to be made in the light of these factors. Persons making evaluations should continually ask: "Do the practices in the school being evaluated accord with the philosophy and objectives of the school and meet the needs of its pupil population and community as well as do the practices of other schools?" They should not consider the size, type, or location of the school, the financial support available, state requirements, or other local factors, except in so far as these factors may have a legitimate effect on the philosophy and objectives of the school or on the needs of the community. In later interpretation of the results of evaluations suitable allowance may be made for any of these factors, but at the time of evaluation an attempt should be made to evaluate the actual program of the school regardless of necessary limitations.

The two-fold nature of the work—evaluation and stimulation to improvement—should also be kept constantly in mind. Careful, discriminating judgment is essential if these purposes are to be satisfactorily served. While the attainment of a high score may be desirable, it is of secondary importance. It should not be permitted to interfere with accurate evaluation; otherwise, real improvement cannot be undertaken and attained.

Those making evaluations should be constantly on guard against the common tendency to choose the higher of two possible evaluations when in doubt. Unless a superior evaluation is definitely indicated and justified by available evidence, one of average or below average should be made.

CHECKLISTS

The checklists consist of provisions, conditions or characteristics found in good secondary schools. Not all of them are necessary, or even desirable, in every good school. Nor do these lists contain all that is desirable in a good school. A school may therefore lack some of the items listed but have other compensating features.

The use of the checklists requires four symbols. (1) If the provision or provisions called for in a given item of the checklist are definitely made or if the conditions indicated are present to a very satisfactory degree, mark the item, in the parenthesis preceding it, with the symbol (+); (2) if the provision is only fairly well made or the conditions are only fairly well met, mark the item with the symbol (−); (3) if the provisions or conditions are needed but are not made, or are very poorly made, or are not present to any significant degree, mark the item with the symbol (0); (4) if it is unnecessary or unwise for the school to have or to supply what specific items call for, mark such items with the symbol (N). (Note: The figures are to be regarded merely as convenient symbols, not mathematical terms.) In brief, mark items:

+ condition or provision is present or made to a very satisfactory degree
− condition or provision is present to some extent or only fairly well made
0 condition or provision is not present or is not satisfactory
N condition or provision does not apply

Space is provided at the end of each checklist for writing in additional items.

EVALUATIONS

Evaluations are to be made, wherever called for, on the basis of personal observation and judgment, in the light of the checklist as marked in accordance with the above instructions, and of all other available evidence, using a five point rating scale, as follows: (Note: The figures are to be regarded merely as convenient symbols, not mathematical quantities.)

5.—*Very superior;* the provisions or conditions are present and functioning to the extent found in approximately the best 10% of regionally-accredited schools.[1]
4.—*Superior;* the provisions or conditions are present and functioning to the extent found in approximately the next 20% of regionally-accredited schools.[1]
3.—*Average;* the provisions or conditions are present and functioning to the extent found in approximately the middle 40% of regionally-accredited schools.[1]
2.—*Inferior;* the provisions or conditions are present and functioning to the extent found in approximately the next 20% of regionally-accredited schools.[1]
1.—*Very inferior;* the provisions or conditions are present and functioning to the extent found in approximately the lowest 10% of regionally-accredited schools.[1]
N.—*Does not apply.* (When this symbol is used, explanation as to the reason the section does not apply should be given under *Comments.*)

Under *Comments* make notations of compensating features or particular shortcomings, explanations, justifications of evaluations, or other pertinent matters.

---

[1] The definitions are given in terms of regionally-accredited schools since the Cooperative Study's experimental program involved primarily regionally-accredited schools. If some other basis of comparison is used, the norms developed from the experimental program will not be applicable.

Figure 54. Instructions for Using the Evaluative Criteria Developed by the Cooperative Study of Secondary School Standards. (From *Evaluative Criteria,* 1940 Edition, Cooperative Study of Secondary School Standards, Washington, page 30.)

Bruner[25] has proposed an elaborate set of criteria for judging courses of study. A gross scale of four points, Excellent, Good, Fair, and Poor, is provided. The following ten questions, for example, are suggested for judging the extent to which the course of study is based upon psychological principles of learning:[26]

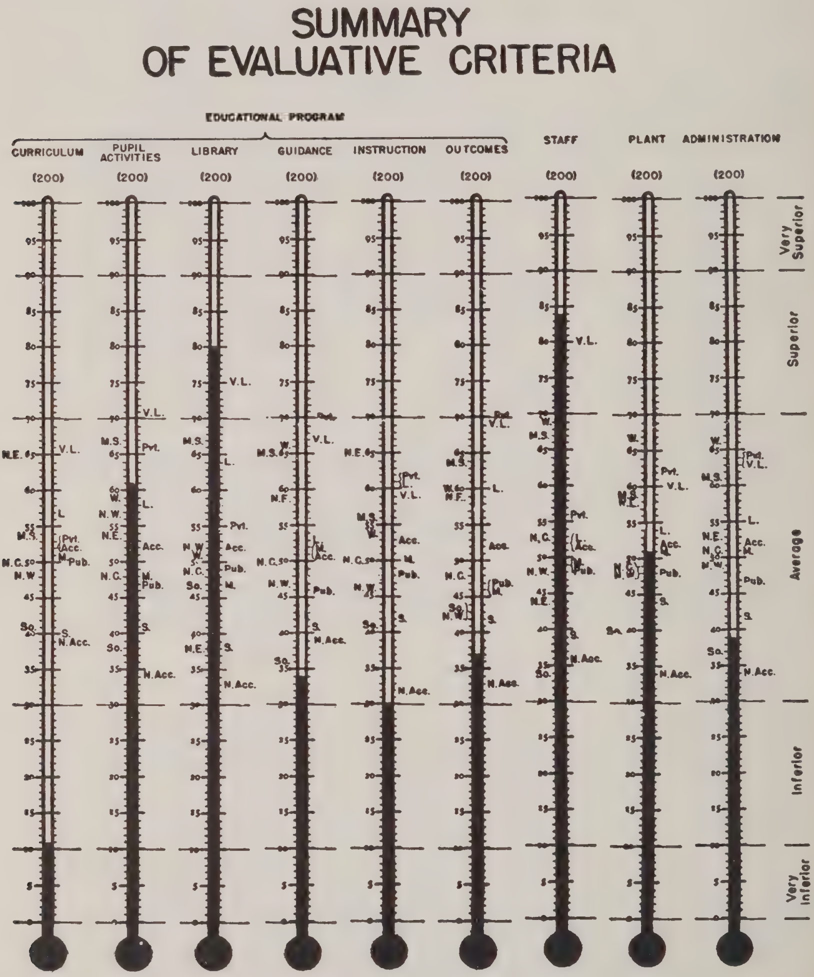


Figure 55. Summary of Evaluative Criteria for the Median Secondary School. (From *How to Evaluate a Secondary School,* 1940 Edition, Cooperative Study of Secondary School Standards, Washington, page 97.)

---

[25] Herbert B. Bruner, "Criteria for Evaluating Course-of-Study Materials," *Teachers College Record,* 39: 107–120, November, 1937.

[26] *Ibid.,* page 111.

---- 1. Is each new learning act considered to be in some degree remaking the whole organism?

---- 2. Is self-activity considered fundamental to learning?

---- 3. Is study conceived of as an attack upon the situation, "and what is learned is learned as and because it is needed for the control of this situation"?

---- 4. Are provisions made for taking into consideration the underlying principles of integration?

B. CONTENT OF OFFERINGS

In evaluating this page consider only content of subject matter offerings, not instructional procedures or methodology. Content should, however, provide not only for informational or factual matter and for skills, but also for understanding the significance of the content and for attitudes, appreciations, and ideals.

A copy of the school's courses of study should be supplied if available; if not, a brief description or outline for each course should be furnished. If the textbook serves as the course of study, it should be evaluated below.

Include in the table only those subjects or courses in which a class is taught every year or in alternate years.

If there are subjects or fields which cannot be classified in the table below, write them in the blank headings or overwrite the headings in one of the columns.

Note that the symbol *N*, "condition or provision does not apply," should be used in the checklist items and evaluations of this table whenever the subject field should not be expected to contribute to the indicated item, or when the subject field is not, and should not be, offered in the school.

CHECKLIST In each major field or area provision is made for:	English	Ancient languages	Modern languages	Mathematics	Sciences	Social studies	Music	Arts and crafts	Industrial arts	Home-making	Agriculture	Business education	Health and phys. educ. for boys	Health and phys. educ. for girls	Vocational shop		
1. Stating the objectives to be attained																	
2. Emphasizing significant contributions of our social heritage to present day life values																	
3. Promoting pupils' understanding of present day social problems																	
4. Stimulating pupils' interests and satisfying their needs																	
5. Modifying courses to meet individual differences																	
6. Including materials and experiences of potential value in adult life																	
7. Interrelating the work in different subject fields																	
8. Suggesting methods to be used in attaining objectives																	
9. Indicating materials to be used or activities to be carried out																	
10. Solving appropriate problems requiring elementary research procedures																	
11. Formulating procedures for evaluating outcomes																	
EVALUATIONS w. *How well does each course of study accord with the philosophy and objectives of the school?*																	
x. *How appropriate is the content of each course of study to meet the needs of the pupil population of the school?*																	
y. *How well is provision made in each course for correlation with other appropriate fields?*																	
z. *How well does each course of study provide for applications to out-of-school life?*																	

Comments:

Figure 56. An Evaluative Procedure for the Content of the Offerings in the Principal Subject-Matter Fields of a Secondary School. (From *Evaluative Criteria*, 1940 Edition, Cooperative Study of Secondary School Standards, page 35.)

____ 5. Are the activities and materials organized into patterns which, if used, assist in the better growing of the individual?

____ 6. Is the position held that the learner should experience satisfaction from engaging in activities?

____ 7. Is knowledge considered as a means to enable the individual to participate more effectively in life situations?

____ 8. Is significance attached to pupil meanings and insights?

____ 9. Is the view held that growth and learning are continuous throughout the life of the individual?

____10. Is provision made for making the situations of the school real and dramatic?

The Cooperative Study suggests a variety of procedures for evaluating the library service of the secondary school. On the assumption that the library service should be a center of the educational life of the school and not merely a collection of books, it is asserted that adequate provisions for the school library should include the following: [27]

> (1) a well educated, efficient librarian; (2) books and periodicals to supply the needs for reference, research, and cultural and inspirational reading; (3) provision for keeping all materials fully cataloged and well organized; (4) a budget which provides adequately for the maintenance and improvement of the library; (5) encouragement of the pupils in the development of the habit of reading and enjoying books and periodicals of good quality and real value.

Figure 57 illustrates the derivation of three measures of the adequacy of the book collection. It will be noted that books of the various classifications are weighted unequally in obtaining the composites. The two extremes are books on philosophy, with a weight of 1, and books on history, travel, and biography, with a weight of 20.

Figure 58 shows the section on Teachers and Libraries and illustrates a different technique. This section seeks answers to two important questions: First, how extensively do the teachers make personal use of the resources of the library in promoting their own professional growth and in their classroom planning and teaching? Second, how effectively do the teachers stimulate pupils to use the library materials?

The Cooperative Study recognizes five areas of guidance responsibility in the secondary school. These are regarded not as distinct types of guidance but rather as phases of an interrelated unitary process. These phases, together with the number of items in the checklists and the evaluations sought, are, in summary, as follows:

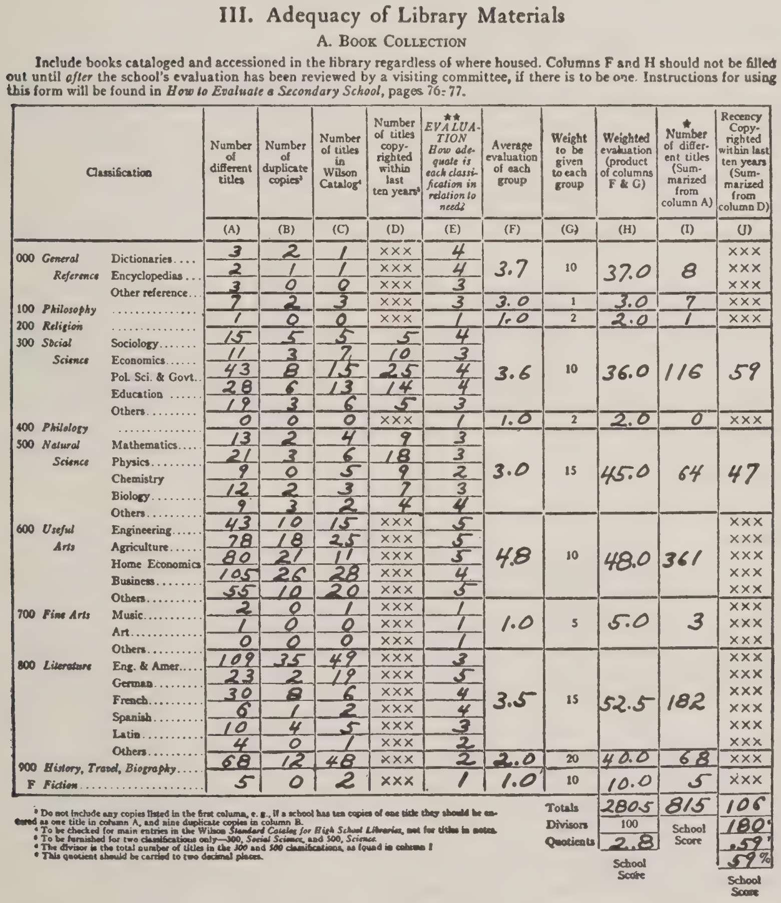
---

[27] *Evaluative Criteria* (1940 Edition), *op. cit.,* page 51.

A. Educational Guidance ....28 items

1. Articulation with lower schools
   How effective are procedures for articulation with lower schools?
2. Curricular and school guidance
   How adequately is guidance provided in such matters as planning a sequence of studies, remedying study difficulties, etc.?
3. Guidance concerning the post-secondary school
   How adequate are provisions for assisting pupils in choices involving the post-secondary school?

B. Vocational Guidance and Placement ....14 items
How adequate are provisions for assisting pupils to make wise vocational choices? How adequate are provisions for placement and follow-up service?



III. Adequacy of Library Materials

A. Book Collection

Include books cataloged and accessioned in the library regardless of where housed. Columns F and H should not be filled out until *after* the school's evaluation has been reviewed by a visiting committee, if there is to be one. Instructions for using this form will be found in *How to Evaluate a Secondary School*, pages 76-77.

Classification		Number of different titles	Number of duplicate copies[3]	Number of titles in Wilson Catalog[4]	Number of titles copyrighted within last ten years[5]	★★ *EVALUATION How adequate is each classification in relation to need?*	Average evaluation of each group	Weight to be given to each group	Weighted evaluation (product of columns F & G)	★ Number of different titles (Summarized from column A)	Recency Copyrighted within last ten years (Summarized from column D)
		(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)
000 *General Reference*	Dictionaries....	3	2	1	×××	4	3.7	10	37.0	8	×××
	Encyclopedias...	2	1	1	×××	4					×××
	Other reference...	3	0	0	×××	3					×××
100 *Philosophy*	....	7	2	3	×××	3	3.0	1	3.0	7	×××
200 *Religion*	....	1	0	0	×××	1	1.0	2	2.0	1	×××
300 *Social Science*	Sociology.......	15	5	5	5	4	3.6	10	36.0	116	59
	Economics......	11	3	7	10	3					
	Pol. Sci. & Govt..	43	8	15	25	4					
	Education	28	6	13	14	4					
	Others.........	19	3	6	5	3					
400 *Philology*	....	0	0	0	×××	1	1.0	2	2.0	0	×××
500 *Natural Science*	Mathematics.....	13	2	4	9	3	3.0	15	45.0	64	47
	Physics.........	21	3	6	18	3					
	Chemistry	9	0	5	9	2					
	Biology.........	12	2	3	7	3					
	Others.........	9	3	2	4	4					
600 *Useful Arts*	Engineering......	43	10	15	×××	5	4.8	10	48.0	361	×××
	Agriculture......	78	18	25	×××	5					×××
	Home Economics	80	21	11	×××	5					×××
	Business.........	105	26	28	×××	4					×××
	Others.........	55	10	20	×××	5					×××
700 *Fine Arts*	Music..........	2	0	1	×××	1	1.0	5	5.0	3	×××
	Art.............	1	0	0	×××	1					×××
	Others.........	0	0	0	×××	1					×××
800 *Literature*	Eng. & Amer.....	109	35	49	×××	3	3.5	15	52.5	182	×××
	German.........	23	2	19	×××	5					×××
	French..........	30	8	6	×××	4					×××
	Spanish.........	6	1	2	×××	4					×××
	Latin...........	10	4	5	×××	3					×××
	Others.........	4	0	1	×××	2					×××
900 *History, Travel, Biography*.....		68	12	48	×××	2	2.0	20	40.0	68	×××
F *Fiction*..................		5	0	2	×××	1	1.0	10	10.0	5	×××
								Totals	2805	815	106
								Divisors	100	School Score	180[6]
								Quotients	2.8		.59[7]
									School Score		59%
											School Score

[3] Do not include any copies listed in the first column, e. g., if a school has ten copies of one title they should be entered as one title in column A, and nine duplicate copies in column B.

[4] To be checked for main entries in the Wilson *Standard Catalog for High School Libraries*, not for titles in notes.

[5] To be furnished for two classifications only—300, *Social Science*, and 500, *Science*.

[6] The divisor is the total number of titles in the 300 and 500 classifications, as found in column I.

[7] This quotient should be carried to two decimal places.

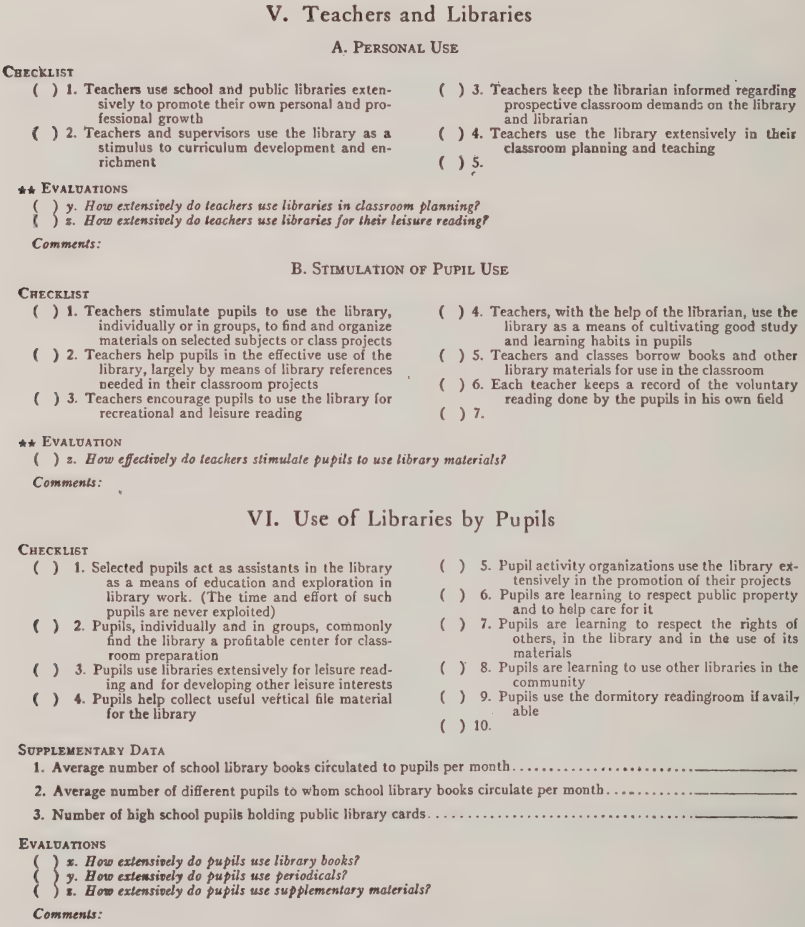
Figure 57. The Computation of Three Measures of the Adequacy of the Book Collections in the Library of a Secondary School. (From *How to Evaluate a Secondary School,* 1940 Edition, Cooperative Study of Secondary School Standards, Washington, page 77.)

C. Guidance in Use of Leisure Time ....6 items
How adequately are pupils assisted in making wise choices of leisure activities?

D. Social and Civic Guidance ....8 items
How adequately are pupils assisted in making wise choices in matters involving social and civic relationships?

E. Personal Guidance ....7 items
How adequately are pupils assisted in making wise choices in personal matters?

Figure 59 illustrates the computation of the summary score for the guidance service of the school when the Alpha Scale is used. It



V. Teachers and Libraries

A. Personal Use

Checklist

( ) 1. Teachers use school and public libraries extensively to promote their own personal and professional growth
( ) 2. Teachers and supervisors use the library as a stimulus to curriculum development and enrichment
( ) 3. Teachers keep the librarian informed regarding prospective classroom demands on the library and librarian
( ) 4. Teachers use the library extensively in their classroom planning and teaching
( ) 5.

**** Evaluations**

( ) y. *How extensively do teachers use libraries in classroom planning?*
( ) z. *How extensively do teachers use libraries for their leisure reading?*

Comments:

B. Stimulation of Pupil Use

Checklist

( ) 1. Teachers stimulate pupils to use the library, individually or in groups, to find and organize materials on selected subjects or class projects
( ) 2. Teachers help pupils in the effective use of the library, largely by means of library references needed in their classroom projects
( ) 3. Teachers encourage pupils to use the library for recreational and leisure reading
( ) 4. Teachers, with the help of the librarian, use the library as a means of cultivating good study and learning habits in pupils
( ) 5. Teachers and classes borrow books and other library materials for use in the classroom
( ) 6. Each teacher keeps a record of the voluntary reading done by the pupils in his own field
( ) 7.

**** Evaluation**

( ) z. *How effectively do teachers stimulate pupils to use library materials?*

Comments:

VI. Use of Libraries by Pupils

Checklist

( ) 1. Selected pupils act as assistants in the library as a means of education and exploration in library work. (The time and effort of such pupils are never exploited)
( ) 2. Pupils, individually and in groups, commonly find the library a profitable center for classroom preparation
( ) 3. Pupils use libraries extensively for leisure reading and for developing other leisure interests
( ) 4. Pupils help collect useful vertical file material for the library
( ) 5. Pupil activity organizations use the library extensively in the promotion of their projects
( ) 6. Pupils are learning to respect public property and to help care for it
( ) 7. Pupils are learning to respect the rights of others, in the library and in the use of its materials
( ) 8. Pupils are learning to use other libraries in the community
( ) 9. Pupils use the dormitory readingroom if available
( ) 10.

Supplementary Data

1. Average number of school library books circulated to pupils per month........................ ____________
2. Average number of different pupils to whom school library books circulate per month............ ____________
3. Number of high school pupils holding public library cards.. ____________

Evaluations

( ) x. *How extensively do pupils use library books?*
( ) y. *How extensively do pupils use periodicals?*
( ) z. *How extensively do pupils use supplementary materials?*

Comments:

Figure 58. Evaluative Techniques for the Library Service of a Secondary School. (From *Evaluative Criteria,* 1940 Edition, Cooperative Study of Secondary School Standards, Washington, page 59.)

will be seen that the various evaluations are entered in spaces provided and then averaged. These point scores are next expressed in percentiles by the use of the standard conversion table. These percentiles are then weighted to obtain a summary score. The equivalent percentile is found from the summary conversion table at the right of the figure. The arrows indicate the sequence of events in the use of the tables. Similar conversion tables are used for the other phases of evaluation.

The quality of instruction in the school is judged by having the work of each member of the teaching staff considered from the following points of view: [28]

A. Classroom Activities.
  1. The teacher's plans and activities.
  2. Cooperation between pupils and teachers.

B. Use of Community and Environment.

C. Textbooks and Other Instructional Materials.
  1. Textbooks.
  2. Other instructional materials.

D. Methods of Appraisal.

E. Special Committee Judgment.

Figure 60 reproduces the last pages of this evaluation and illustrates the procedure.

The philosophy underlying the evaluation of the outcomes of the educational program is clearly stated in the following guiding principles: [29]

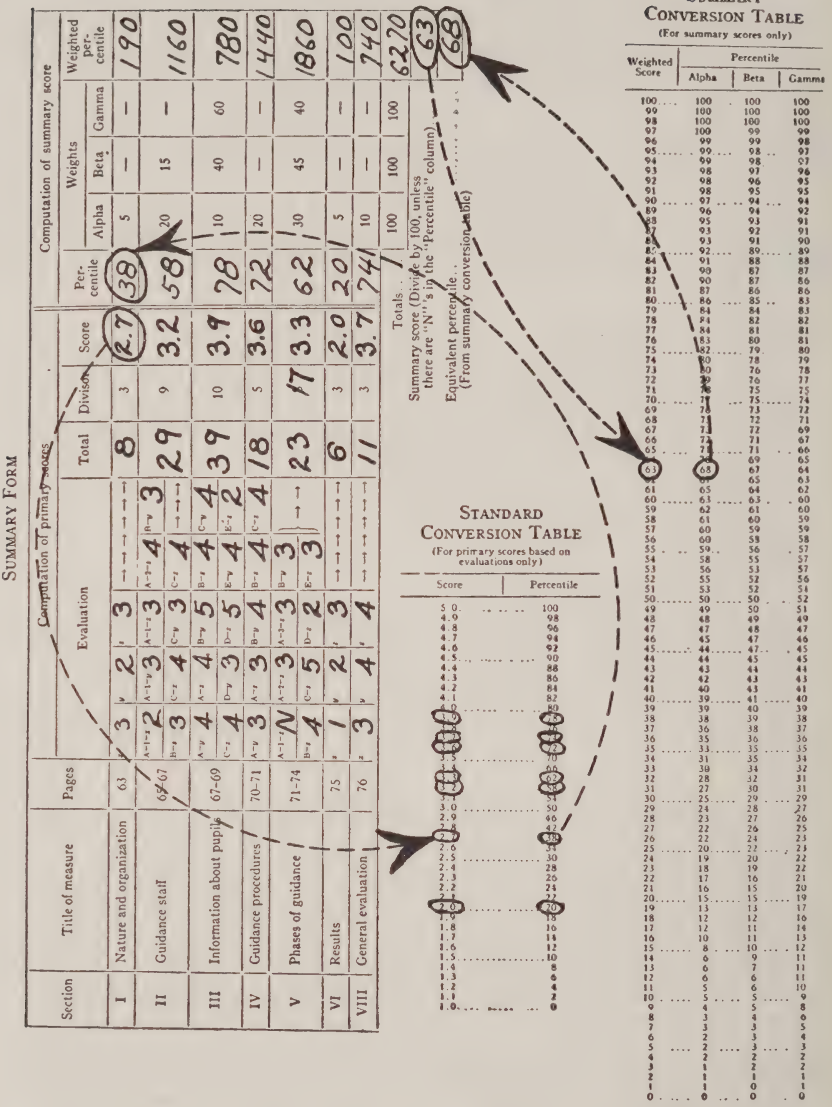
> In the educational program of a good secondary school, major concern should be given to attaining desirable outcomes and to the various kinds of evidence indicating that such outcomes are being realized. It may be necessary to test some outcomes by departments or in class groups. This, however, should not be construed as limiting the responsibilities of all phases of the educational program, including the instructional activities of teachers, pupil activity program, guidance service, library service, school plant, and school administration, for the attainment of desirable outcomes. There should be evidence that teachers and pupils are happily and harmoniously cooperating in the stimulation of a wholesome curiosity about themselves and their environment. Evidence should be sought to show that pupils are securing knowledge and developing worthwhile skills, attitudes, tastes, appreciations, and habits. There should be evidence that pupils are able to make desirable choices or to exercise good judgment in the selection of friends, vocations, leisure activities, goods and services, and in other important matters which confront youth today. Evaluation of such activities involves more than determining the amount of knowledge possessed, measuring the degree of skill, and testing the scope of understanding, important and necessary as all these are.

---

[28] *Evaluative Criteria* (1940 Edition), *op. cit.,* page 160.

[29] *Op. cit.,* page 83.

V. Guidance Service



SUMMARY FORM

			Computation of primary scores							Computation of summary score			
										Weights			
Section	Title of measure	Pages	Evaluation		Total	Divisor	Score	Per-centile	Alpha	Beta	Gamma	Weighted per-centile
I	Nature and organization	63	x 3, y 2, z 3	→ → → →	8	3	2.7	38	5	—	—	190
II	Guidance staff	64-67	A-1-x 2, A-1-y 3, A-1-z 3, A-2-x 4, B-y 3	B-x 3, C-x 4, C-y 3, C-z 4, → →	29	9	3.2	58	20	15	—	1160
III	Information about pupils	67-69	A-y 4, A-z 4, B-y 5, B-z 4, C-y 4	C-z 4, D-y 3, D-z 5, E-y 4, E-z 2	39	10	3.9	78	10	40	60	780
IV	Guidance procedures	70-71	A-y 3, A-z 3, B-y 4, B-z 4, C-z 4		18	5	3.6	72	20	—	—	1440
V	Phases of guidance	71-74	A-1-z N, A-2-z 3, A-3-z 3, B-y 3, → →	B-z 4, C-z 5, D-z 2, E-z 3	23	7	3.3	62	30	45	40	1860
VI	Results	75	x 1, y 2, z 3	→ → → →	6	3	2.0	20	5	—	—	100
VIII	General evaluation	76	x 3, y 4, z 4	→ → → →	11	3	3.7	74	10	—	—	740
								Totals	100	100	100	6270
								Summary score (Divide by 100, unless there are "N"'s in the "Percentile" column)				63
								Equivalent percentile (From summary conversion table)				68

STANDARD CONVERSION TABLE

(For primary scores based on evaluations only)

Score	Percentile
5.0	100
4.9	98
4.8	96
4.7	94
4.6	92
4.5	90
4.4	88
4.3	86
4.2	84
4.1	82
4.0	80
3.9	78
3.8	76
3.7	74
3.6	72
3.5	70
3.4	66
3.3	62
3.2	58
3.1	54
3.0	50
2.9	46
2.8	42
2.7	38
2.6	34
2.5	30
2.4	28
2.3	26
2.2	24
2.1	22
2.0	20
1.9	18
1.8	16
1.7	14
1.6	12
1.5	10
1.4	8
1.3	6
1.2	4
1.1	2
1.0	0

SUMMARY CONVERSION TABLE

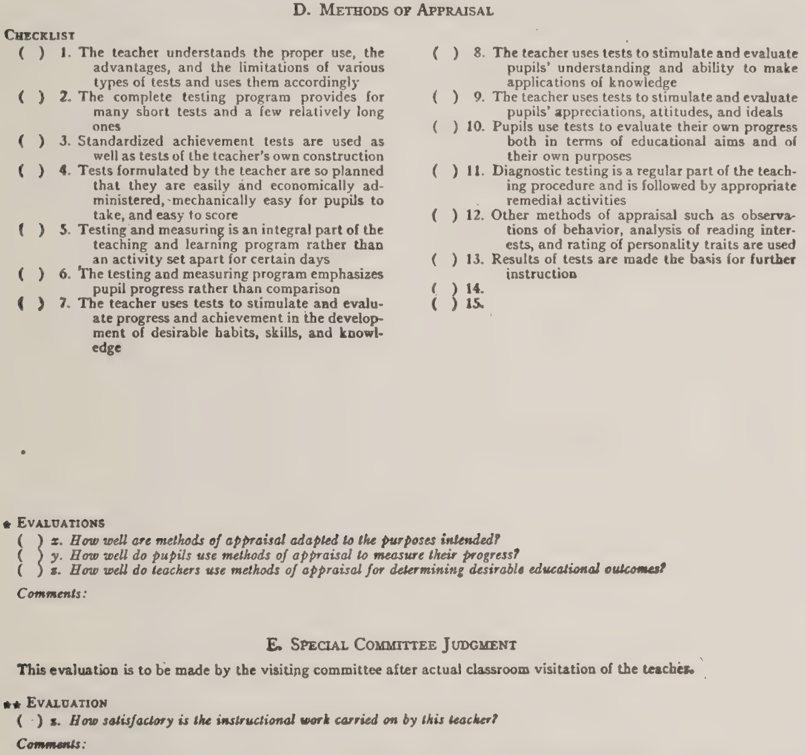
(For summary scores only)

Weighted Score	Percentile		
	Alpha	Beta	Gamma
100	100	100	100
99	100	100	100
98	100	100	100
97	100	99	99
96	99	99	98
95	99	98	97
94	99	98	97
93	98	97	96
92	98	96	95
91	98	95	95
90	97	94	94
89	96	94	92
88	95	93	91
87	93	92	91
86	93	91	90
85	92	89	89
84	91	88	88
83	90	87	87
82	90	87	86
81	87	86	86
80	86	85	83
79	84	84	83
78	84	82	82
77	84	81	81
76	83	80	81
75	82	79	80
74	80	78	79
73	80	76	78
72	79	76	77
71	78	75	75
70	77	75	74
69	76	73	72
68	73	72	71
67	73	72	69
66	72	71	67
65	71	71	66
64	70	69	65
63	68	67	64
62	67	65	63
61	65	64	62
60	63	63	60
59	62	61	60
58	61	60	59
57	60	59	59
56	60	58	58
55	59	56	57
54	58	55	57
53	56	53	57
52	55	52	56
51	53	52	54
50	50	50	52
49	49	50	51
48	48	49	49
47	47	48	47
46	45	47	46
45	44	47	45
44	44	45	45
43	43	44	44
42	42	43	43
41	40	43	41
40	39	41	40
39	39	40	39
38	38	39	38
37	36	38	37
36	35	36	36
35	33	35	35
34	31	35	34
33	30	34	32
32	28	32	31
31	27	30	31
30	25	29	29
29	24	28	27
28	23	27	26
27	22	26	25
26	22	24	23
25	20	22	23
24	19	20	22
23	18	19	22
22	17	16	21
21	16	15	20
20	15	15	19
19	13	13	17
18	12	12	16
17	12	11	14
16	10	11	13
15	8	10	12
14	6	9	11
13	6	7	11
12	6	6	11
11	5	6	10
10	5	5	9
9	4	5	8
8	3	4	6
7	3	3	5
6	2	3	4
5	2	3	3
4	2	2	2
3	1	2	2
2	1	1	1
1	1	0	1
0	0	0	0

Figure 59. The Computation of the Summary Score for the Guidance Service of a Secondary School. (From *How to Evaluate a Secondary School*, 1940 Edition, Cooperative Study of Secondary School Standards, Washington, pages 82 and 86.)

Among others, intangible qualities such as cooperativeness, tolerance, open-mindedness, reverence, respect for law, and self-reliance are highly desirable outcomes. Evaluation of such outcomes is by no means easy; for most of them there is no standard measure and therefore evaluation of them necessarily will be largely a matter of judgment. The difficulty of the task is no reason for avoiding it, and the importance and universality of the problems involved make it imperative that attention should be directed to the attainment of such outcomes and to their proper evaluation.

Another useful instrument "designed to serve as a basis for the appraisal of individual school systems with respect to their adaptation to current educational needs" has been prepared by Mort and Cornell.[30] It covers much the same scope as the Cooperative Study, but the technique is different. Specific questions are raised, to be



D. METHODS OF APPRAISAL

CHECKLIST

( ) 1. The teacher understands the proper use, the advantages, and the limitations of various types of tests and uses them accordingly
( ) 2. The complete testing program provides for many short tests and a few relatively long ones
( ) 3. Standardized achievement tests are used as well as tests of the teacher's own construction
( ) 4. Tests formulated by the teacher are so planned that they are easily and economically administered, mechanically easy for pupils to take, and easy to score
( ) 5. Testing and measuring is an integral part of the teaching and learning program rather than an activity set apart for certain days
( ) 6. The testing and measuring program emphasizes pupil progress rather than comparison
( ) 7. The teacher uses tests to stimulate and evaluate progress and achievement in the development of desirable habits, skills, and knowledge
( ) 8. The teacher uses tests to stimulate and evaluate pupils' understanding and ability to make applications of knowledge
( ) 9. The teacher uses tests to stimulate and evaluate pupils' appreciations, attitudes, and ideals
( ) 10. Pupils use tests to evaluate their own progress both in terms of educational aims and of their own purposes
( ) 11. Diagnostic testing is a regular part of the teaching procedure and is followed by appropriate remedial activities
( ) 12. Other methods of appraisal such as observations of behavior, analysis of reading interests, and rating of personality traits are used
( ) 13. Results of tests are made the basis for further instruction
( ) 14.
( ) 15.

* EVALUATIONS
( ) *x. How well are methods of appraisal adapted to the purposes intended?*
( ) *y. How well do pupils use methods of appraisal to measure their progress?*
( ) *z. How well do teachers use methods of appraisal for determining desirable educational outcomes?*

Comments:

E. SPECIAL COMMITTEE JUDGMENT

This evaluation is to be made by the visiting committee after actual classroom visitation of the teacher.

** EVALUATION
( ) *z. How satisfactory is the instructional work carried on by this teacher?*

Comments:

Figure 60. An Evaluative Technique for the Quality of Instruction in a Secondary School. (From *Evaluative Criteria*, 1940 Edition, Cooperative Study of Secondary School Standards, Washington, page 160.)

---

[30] Paul R. Mort and Francis G. Cornell, *A Guide for Self-Appraisal of School Systems*, 59 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1937.

answered *Yes* or *No,* with places for the supporting data at the left of the page. The scores for each section are then entered on a special score sheet, and by a simple process of weighting are combined into a single score. Table 51 gives the summary of this score sheet. Tentative norms are available for school systems of various sizes located in four states. The first of the ten parts of the section, which attempts to determine the degree to which the educational program recognizes the nature and extent of individual differences in pupils, is as follows: [31]

a. Intelligence Tests. Group and individual intelligence tests should be used as one of the means of analyzing problems of maladjustment.

Q. *How many of your pupils have been given intelligence tests?*

Interview: Principal, Guidance director, Psychologist (if any).
Observe: Individual records, Test records.

Evidence ....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....

1. Individual intelligence tests have been given in special cases. Yes... No...
2. A group intelligence test is given to all elementary and first year high school pupils at least once in three years. Yes... No...
3. Intelligence tests results, for tests given both in elementary and in high school, are made a part of the permanent record of the child. Yes... No...
4. Educational and intelligence tests results (group and individual) constitute one of the elements upon which guidance is based. Yes... No...

The school organization and plant. Numerous checklists and score cards have been prepared for rating school buildings, equipment, and administrative procedures. In this field, the pioneer work of Strayer and Engelhardt [32] and their students is especially notable. Two recent developments will be described briefly. The scope of the evaluation procedures devised by the Cooperative Study is apparent from the following outlines of evaluative criteria:

---

[31] *Ibid.*, page 26.

[32] Published by Bureau of Publications, Teachers College, Columbia University, New York.

A. School Staff:
  1. Numerical adequacy.
  2. Professional staff: selection, qualifications, improvement.
  3. Nonprofessional staffs: qualifications, improvement in and conditions of service.
  4. Special characteristics of the school staff.
  5. General evaluation.

B. School Plant:
  1. The site: health and safety, economy and efficiency, influence on the educational program.
  2. The building: health and safety, economy and efficiency, influence on the educational program.
  3. Equipment: health and safety, economy and efficiency, influence on the educational program.
  4. Special services: cafeterias, clinics, etc.
  5. Special characteristics of the school plant.
  6. General evaluation.

C. School Administration:
  1. Administrative staff: numerical adequacy, preparation and qualifications, improvement in service.
  2. Organization: board of control, general policies, superintendent of schools, principal.
  3. Supervision of instruction: objectives, procedures and activities, principles, results.
  4. Supervision of special services.
  5. Business management: general duties and procedure, budget, accounting, maintenance and operation.
  6. School and community relations.
  7. Special characteristics of the school administration.
  8. General evaluation.

That the scope of the analysis proposed by Mort and Cornell is somewhat similar is apparent from Table 51. Short sections relating to the school site will illustrate the differences in the two techniques.

B. ECONOMY AND EFFICIENCY[33]

CHECKLIST
( ) 1. The site is readily accessible to the school population.
( ) 2. It is accessible over hard surfaced roads and adequate walks.
( ) 3. It is sufficiently extensive for building and play needs, driveways, and landscaping.
( ) 4. Play areas are readily accessible.
( ) 5. The site has possibility of future expansion, extension, or adaptation without too great cost.

---

[33] *Evaluative Criteria* (1940 Edition), *op. cit.,* page 116.

Table 51

SUMMARY OF THE MORT-CORNELL SCORE SHEET FOR THE SELF-APPRAISAL OF SCHOOL SYSTEMS

Section	Adjustments Possible		Maximum Score	
	Section	*Total*	*Section*	*Total*
I. Classroom Instruction:				
A. The curriculum		30		210
1. Flexibility of curriculum	10		70	
2. Breadth of curriculum	10		70	
3. Courses of study	10		70	
B. Pupil activity		28		196
1. Fields of learning	13		91	
2. Extracurricular activities	7		49	
3. Instructional materials	8		56	
II. Special Services for Individual Pupils:				
A. Pupil records and attendance		13		104
1. Educational accounting	7		56	
2. Census and attendance	6		48	
B. Provisions for individual differences:		26		156
1. Guidance: educational and vocational	7		42	
2. The individual and the educational program	10		60	
3. Health service	9		54	
III. Educational Leadership:				
A. Supervision and school organization		21		105
1. Professionalization of personnel	8		40	
2. Supervision of instruction	8		40	
3. Grade and subject organization	5		25	
B. School administration and the community		21		105
1. Administrative planning	6		30	
2. Status of control	7		35	
3. Scope of school influence in the community	8		40	
IV. Physical Facilities and Business Management:				
A. The school plant		30		90
1. School plant planning	5		15	
2. The school site	5		15	
3. School buildings	10		30	
4. Special rooms	10		30	
B. Business management		14		42
1. Supplies and equipment	7		21	
2. Financial accounting	7		21	
Total		183		1,008

( ) 6. It is as near the center of the school population as environmental conditions make advisable.

( ) 7.

( ) 8.

Evaluations

( ) *x. How accessible is the site?*

( ) *y. How extensive is the site?*

( ) *z. How well adapted is the site for future expansion?*

Comments:

THE SCHOOL SITE[34]

d. Adaptability. Each school site should be laid out and developed in consideration of both present and estimated future needs. YES... NO...

Q. *What is the average size of the elementary school sites? Of high school sites? How do you justify the size?*

Interview: Superintendent.

Observe: Plans and data on future growth, areas of present sites and enrollment of schools.

Evidence ....
....
....
....
....
....
....
....
....
....

1. The superintendent has developed plans for the layout of each permanent site in terms of estimated changes in enrollment and educational program. Yes... No...
2. Present development of sites is such that adjustments can be made with minimum of cost to expanding enrollment and program. Yes... No...

An index of variation. N. L. Engelhardt, Jr.,[35] has recently prepared an *index of variation,* which is based upon the assumption that any true equalization of educational opportunities must provide for variability, rather than uniformity, in the school plant and program. The needed variation must consider the impact upon the pupil of the physical and social characteristics of the environment, as well as the personal qualities of the people to be served. For example, the thirty-eight factors that should be taken into account in determining the educational program of a community relate to the age distribution of the population, the health conditions, the housing conditions, and the social conditions of the community.

---

[34] *A Guide for Self-Appraisal of School Systems, op. cit.,* page 49.

[35] *The Report of A Survey of the Public Schools of Pittsburgh, Pennsylvania,* pages 437–444. New York: Bureau of Publications, Teachers College, Columbia University, 1940. Also see *The American School and University Yearbook* for 1940.

Concluding statement. Evaluation is by no means a new idea in education, although the concept has been greatly enlarged in recent years. Many new techniques have been devised to supplement those already in existence, and in some cases to supplant them altogether. Much remains to be done, however. In the meantime educators should acquaint themselves with the uses and limitations of the techniques which have been developed. There is no escaping the fact that evaluation is one of the most difficult, as well as one of the most important, problems in the modern school. The best existing evidence that a school is good is the fact that it is continually studying to find ways to improve itself.

SELECTED REFERENCES FOR FURTHER STUDY

"Appraising the Elementary-School Program," *The National Elementary Principal,* 16: 227–655, July, 1937.

Evaluation of Secondary Schools: General Report. Washington, D. C.: Cooperative Study of Secondary School Standards, 1939. 526 pages.

Evaluative Criteria (1940) Edition). Washington, D. C.: Cooperative Study of Secondary School Standards, 1939. 176 pages.

How to Evaluate a Secondary School (1940 Edition). Washington, D. C.: Cooperative Study of Secondary School Standards, 1939. 139 pages.

Leonard, J. Paul, and Eurich, Alvin C., Editors, *An Evaluation of Modern Education.* New York: D. Appleton-Century Company, 1942. 299 pages.

Mort, Paul R., and Cornell, Francis G., *A Guide for Self-Appraisal of School Systems.* New York: Bureau of Publications, Teachers College, Columbia University, 1937. 59 pages.

Reavis, W. C., and Cooper, D. H., *Evaluation of Teacher Merit in City School Systems.* Chicago: University of Chicago Press, 1945. 139 pages.

Troyer, Maurice E., and Pace, C. Robert, *Evaluation in Teacher Education.* Washington, D. C., American Council on Education, 1944. 368 pages.

Zook, George F., and Haggerty, M. E., *Principles of Accrediting Higher Institutions.* Chicago: University of Chicago Press, 1934. 202 pages.

CHAPTER XVIII

Public Relations

A. The Problem

> Much of the scepticism prevalent as to the power and value of popular education arises from the inability of the educationist, or of the school teacher, to adduce satisfactory statistical evidence of the moral or of the intellectual results from any special courses of instruction or training, as manifested in after life.[1]

Thus began an article pointing out the need for better measurement in education, by E. Chadwick, published in an English educational journal in 1864. Seventy-five years later a survey of parent opinion of public education in America revealed deep concern over "teacher-made courses of study" in which the parents have been allowed no voice, as well as concern over "the stress schools place upon the spectacular in education to the detriment of programs that will improve the manners and morals of children."[2] These statements indicate that the public and the school do not always understand each other, and consequently do not work together to their mutual advantage.

The meaning of public relations programs. Narrowly conceived, the public relations program of the school is synonymous with the publicity activities of the school. In recent years, however, the terms *publicity* and *propaganda* have become so closely associated and so discredited in the public mind as to arouse suspicion that something sinister is about to be "put over." Broadly conceived, *public relations* is merely one important aspect of the school's program of adult education. Its primary aims are two: (1) better understanding by the public of the purposes, programs, accomplishments, and needs of the school, which, in the words of Commissioner Cook, is "both a professional opportunity and a professional obligation"; and (2) better understanding by the school of the desires and needs of the community as reflected in the educational views of the public. In other words, its purpose is to effect the maximum co-operation between the community's two most important educational institutions, the home and the school. And it must be remembered always that the child is the connecting link between them.

---

[1] Quoted by Edward L. Thorndike in *Journal of Educational Psychology,* 4: 551, October, 1913.

[2] Lester S. Ivins, "What Parents Expect of the School," *Journal of the National Education Association,* 28: 194, October, 1939.

A prominent educational leader[3] includes among the important purposes of measurement and evaluation in the modern school the provision of "psychological security to the school staff, to the pupils, and to the parents," and a "sound basis of public relations." Concerning the latter Tyler says:

> No factor is so important in establishing constructive and co-operative relations with the community as an understanding on the part of the community of the effectiveness of the school. A careful and comprehensive evaluation should provide evidence that can be widely publicized and used to inform the school community about the value of the school program. Many of the criticisms of the school expressed by the taxpayers and parents can be met and turned into constructive co-operation if concrete evidence is available regarding the accomplishments of the school.

There are several reasons for thinking that the problem is becoming increasingly important and difficult as the years go by. The enlarged enrollment in the secondary school and the accompanying expansion in the school program have brought many changes which the public does not understand. This fact is mainly responsible for the common charge that the modern school curriculum is cluttered up with all sorts of useless "fads and frills." The increasing burden of taxation has naturally made the citizens critical of all public expenditures. Since in most communities the public school system is the biggest public business, it is likely to bear the brunt of the attack. Nor should one overlook the stubborn fact that the enormous expansion of such enterprises as are provided for by the social security and old-age pension legislation has greatly increased the competition for the taxpayer's dollar. In such a situation it is especially well to keep in mind the wise statement of President Madison: "A popular government without popular information or the means of acquiring it, is but a prologue to a farce or a tragedy, or perhaps both."[4]

The principal sources of "popular information" may be conveniently grouped as follows:

1. Ordinary agencies: local newspapers, student publications.
2. Official publications: reports, bulletins, handbooks, etc.

---

[3] Ralph W. Tyler, "The Place of Evaluation in Modern Education," *Elementary School Journal,* 41: 19–27, September, 1940.

[4] For an excellent statement of the philosophy underlying an effective program of public-school relations, see: Ward G. Reeder, *An Introduction to Public-School Relations,* Chapter I. New York: The Macmillan Company, 1937.

3. Report cards and letters to parents.
4. Miscellaneous: public programs, exhibits, P.T.A., etc.

Each of these will now receive brief discussion.

B. Ordinary Agencies of Public Information

Local newspapers. As a medium for bringing about desirable relations between the school and its public, the newspaper ranks high. For most people it is the principal source of information, but school news as reported in the local paper is likely to be narrow in scope and lack proportion. An extensive study by Farley[5] revealed that as a rule the patrons received least information on the school topics in which they were most interested and most information on school topics in which they were least interested. Table 52 summarizes the situation.[6]

TABLE 52

THE RANK ORDERS OF THIRTEEN TOPICS OF SCHOOL NEWS ACCORDING TO THE INTERESTS OF 5,067 SCHOOL PATRONS COMPARED WITH THE SPACE DEVOTED TO THESE TOPICS BY TEN NEWSPAPERS (AFTER FARLEY)

Topics of School News	Rank According to	
	Patrons' Interests	*Space in News*
Pupil progress and achievement	1	4
Method of instruction	2	10
Health of pupils	3	9
Courses of study	4	6
Value of education	5	12
Discipline and behavior of pupils	6	11
Teachers and school officers	7	2
Attendance	8	13
Buildings and building program	9	8
Business management and finance	10	7
Board of education and administration	11	5
Parent-teacher association	12	3
Extracurricular activities	13	1

It may not be surprising but it is certainly unfortunate to find that in the typical newspaper the total space devoted to the first six items in order of patrons' interests is less than half that given to extracurricular activities, which stand at the bottom of the list. Both the school and newspaper appear to take for granted the excellent work of the classroom, which, therefore, falls in the dog-bites-man category rather than in the news classification.[7] They both apparently forget that a report of the incident is the most interesting thing in the world to the owner of the dog as well as to the man who has been bitten. Parents never tire of hearing good reports of their own children. There is no good reason why the educational side shows should be allowed to swallow up the main tent. Farley calls attention to these facts:[8]

---

[5] Belmont Mercer Farley, *What to Tell the People about the Public Schools,* 136 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1929.

[6] Adapted from pages 16 and 49.

> In other words, patrons wish to know *what* their children are being taught, *how* they are being taught, *what results* are being achieved, and how the public schools affect the physical welfare of their children. . . . They are ready to listen to the educator tell them that the results achieved in the schools are desirable, that they are achieved by efficient, scientific methods, that children are taught useful habits and skills, that their physical welfare is not neglected.

Student publications. Student publications should occupy a strategic position in any public relations program. They represent activities that have educational value in themselves and thus constitute important exhibits of the actual work of the school. Of these publications the school newspaper and the yearbook or annual are most important. Since they are written primarily for the pupils and patrons of the school, these publications can portray the actual operation of the school program more fully than the general newspaper, which must appeal to a wider public. What the student does is always of interest to other students and to parents but examination of the student publications of most schools would probably reveal a very distorted picture of the school situation. As in the regular newspaper, the extracurricular program looms large. The reader can scarcely escape the conclusion that the school year is largely occupied with social affairs and athletics. Those who criticize public education as an exponent of "fads and frills" could hardly do better than introduce the yearbook as Exhibit A. Beside the stadium, the library dwindles into insignificance, and such things as classrooms and laboratories are deemed so unimportant as to be omitted altogether. It is not too much to expect that the student publications present a truer picture of the school, giving greater prominence to those features which justify its existence. That the public is genuinely interested in these, there can be no doubt. Certainly parents would put evidence of pupil progress and achievement at the top of the list.

---

[7] For an instructive discussion of this point, see: Edwin J. Brown, *Secondary-School Administration,* pages 270–271. Boston: Houghton Mifflin Company, 1938.

[8] *Ibid.,* pages 16, 17.

C. Official Publications

Annual reports. The earliest record of a formal written educational report was made in Boston, Massachusetts, in 1738, although informal oral reports had been made to town meetings in New England at an earlier period.[9] It is clear that from the outset the primary function of such reports has been to inform the public regarding the aims, progress, and needs of the schools, and to afford an intelligent basis for determining educational policies. The first written report, for example, gave the enrollment in each school and included comments by the visiting committee on the quality of instruction. The function of such reports was well stated in the introductory pages of the 1841–1842 report of Fall River, Massachusetts, as follows:[10]

> Those who are taxed to support Public Schools, have a right to know how their money is expended, and what is the character of the schools which they are required to maintain. The committee are but the agents employed by the town to take the agency of Common School Education, and the employer ought to be made acquainted with all that appertains to his interest, in respect to this agency. What the committee knows as to the schools, the town ought to know.

Since the appearance of standardized tests, the annual reports often describe the tests used and the purposes for which they are employed, and give summaries of the results. Some cities, Detroit and Cleveland,[11] for instance, make effective use of graphs to show that progress in the tool subjects is regular from grade to grade, as well as profile charts to illustrate the use of standard tests in the diagnosis and guidance of individual pupils. There is no way better than test results to show the need for curriculum changes, guidance services, ungraded classes, and other provisions for individual differences. There can be little doubt that parents are interested in receiving not only an account of how the money for public education was spent, but also of what it bought in the way of an efficient educational program.

But most school reports have one fatal weakness: They are not read. The reason for this has been clearly stated as follows: "Most official reports are dull. Their authors, though they have the most interesting material in the world, treat it perfunctorily, statistically, as lifeless stuff to be put away in mortuary files."[12] The problem with school reports, as G. Stanley Hall long ago pointed out in the case of moral education, is how to make virtue exciting.[13]

---

[9] Ward G. Reeder, *op. cit.*, pages 85–87.

[10] Quoted from M. G. Neale, *School Reports as a Means of Securing Additional Support for Education in American Cities,* pages 4–5. Columbia, Mo.: Missouri Book Company, 1921.

[11] Clifford Woody and Paul V. Sangren, *Administration of the Testing Program,* Chapter IX. Yonkers: World Book Company, 1933.

[12] Editorial in *The New York Times,* January 4, 1926.

Special reports and publications. It must be recognized that nothing is great or small, good or bad, except by comparison. Because of this fact, school surveys, which attempt to interpret the local schools in relation to those of other systems of similar size, are important. At times, such studies made by impartial outside agencies are especially effective. It is even better, perhaps, to have a continuous self-survey, and to report at strategic intervals various phases of the school program. The larger cities employ for this purpose bulletins or magazines modeled after the house organs of industrial organizations. Graphical comparisons of standardized test scores with national norms may be so reported.

A common criticism of the modern school program is that it has allowed the newer "fads and frills" to displace the older "fundamental subjects." People long for "the good old days when people really learned something when they went to school." The most effective argument with which to meet such criticism is a comparison of the achievement of the older schools and the newer, or of the traditional school program and the more liberal program of today. Riley[14] made such a study in Springfield, Massachusetts, of the results of tests in 1906 that had first been given to children in the city sixty years earlier, in 1846. The results, briefly summarized below in terms of percentage of correct responses were favorable to the later schools.

SUBJECTS	PERCENTAGE CORRECT	
	1846	*1905–1906*
Arithmetic	29.4	65.2
Spelling	40.6	51.2
Geography	40.3	53.4

Fish[15] made a somewhat similar study comparing the achievement of Boston children in 1928 with that of pupils in the city on the same tests in 1853, seventy-five years earlier. Again the results, expressed in terms of errors made, favored the later schools:

SUBJECTS	ERRORS MADE	
	1853	*1928*
Arithmetic	5.4	1.6
Grammar	6.5	3.1
Geography	4.4	4.2

---

[13] For a good discussion see: Ward G. Reeder, *op. cit.*, pages 89–104.

[14] J. L. Riley, *The Springfield Tests.* Springfield, Mass.: The Holden Patent Book Company, 1908.

[15] Louis J. Fish, *Examinations Seventy-Five Years Ago and Today.* Yonkers: World Book Company, 1930.

These studies suggest preserving the results of standardized tests so that at intervals of perhaps ten or twenty years they can be compared with current results on these tests, which will afford convincing evidence of trends in efficiency. A recent study of this type, covering achievement in Philadelphia high schools for a ten-year period, has been made by Boyer and Gordon;[16] and a study of arithmetic for a twelve-year period in St. Louis has been made by Boss.[17]

D. Report Cards and Letters to Parents

Trends in report cards. For many years report cards have furnished the most direct line of communication between the home and the school. They have ordinarily consisted of a record of the pupil's attendance and academic achievement, expressed in teachers' marks, sent to the parent at intervals of a month or six weeks. In recent years, however, certain important changes have taken place. In a comprehensive survey of the literature relating to report cards, Messenger and Watts[18] note the following trends:

1. There is general dissatisfaction with any scheme of grading that encourages the comparison of pupils with each other.
2. If any grades are used, a scale with fewer points is favored, a three-point scale being most often recommended.
3. There is a wide-spread feeling that the schools should evaluate traits other than mere subject-matter achievement.
4. There is a clear tendency to use descriptive rather than quantitative reports.
5. Report cards are being displaced by notes or letters to parents.
6. Cards, notes, or letters are being sent at less frequent intervals and in some schools only when there is specific occasion for such communications.
7. Attempts are being made to give more detailed diagnosis of pupils' achievements.
8. Parents are being asked to cooperate in building report forms.
9. Pupils are cooperating both in devising report cards and in evaluating their own accomplishment.

A more recent study[19] of trends in nine western states indicates that these changes have been more marked in the elementary than in the secondary school, as is evident from the following frequencies reported: primary, 88; intermediate, 78; upper grades, 54; junior high, 38; and senior high school, 23. This study notes a wholesome effect on the personalities of the pupils, the effect being especially marked for those of lower ability. The most noticeable effect, however, appears to be in improved teacher-pupil relationships.

---

[16] Philip A. Boyer and Hans C. Gordon, "Have High Schools Neglected Academic Achievement?", *School and Society,* 49: 810–812, June 24, 1939.

[17] Mabel E. Boss, "Arithmetic, Then and Now," *School and Society,* 51: 391–392, March 23, 1940.

[18] Helen R. Messenger and Winifred Watts, "Summaries of Selected Articles on School Report Cards," *Educational Administration and Supervision,* 21: 539–550, October, 1936.

[19] Henry H. Hartley, "Report Card Trends in West," *Nation's Schools,* 24: 51–53, November, 1939.

There is also abundant evidence that wherever these newer systems of reporting have been adopted, they have received practically the unanimous approval of the parents. After six years' experience, for example, one writer makes this positive statement: "The letter fosters a much more co-operative relation between home and school." [20] Morrisett [21] reports a study in which the principal of a large junior high school submitted a list of forty items to the parents with the instruction to check "items in which you are most interested; that is, those items about which you would like to know more." The item "What parents can do to promote pupil accomplishment" ranked first. Other items high in the list clearly indicated that parents desired more information regarding educational and vocational guidance. The weakness of the older report card was just here. The information supplied to parents, even if its accuracy could be assumed, was of such a general character as to be of little help in either diagnosis or guidance, in which full co-operation with the home is most needed.

Evans [22] has traced the evolution of the report card. He notes a definite trend away from the standardized printed card and toward a more flexible, informal report that is better adapted to local conditions and needs. There is an increasingly clear recognition that the function of reporting is *interpretation* rather than presentation, with the emphasis on *progress* rather than on status.

Hill's study of report cards. Hill [23] analyzed 443 report cards from towns and cities of all sizes, representing all educational levels and practically every state. He concluded that a satisfactory report card should:

> 1. Represent the true spirit, purposes, and functions of the school. . . .
> 2. Reflect educational objectives arrived at only after careful consideration and mature judgment.
> 3. Change in accord with changes in educational standards and educational philosophy. . . .

---

[20] V. L. Beggs, "Reporting Pupil Progress without Report Cards," *Elementary School Journal,* 37: 107–114, October, 1936.

[21] L. N. Morrisett, "Interpreting the School to the Public," *Clearing House,* 7: 480–485, April, 1933.

[22] Robert O. Evans, *Practices, Trends, and Issues in Reporting to Parents on the Welfare of the Child in School,* 98 pages. New York: Bureau of Publications, Teachers College, Columbia University, 1938.

[23] George E. Hill, "The Report Card in Present Practices," *Educational Method,* 15: 115–131, December, 1935.

4. Present a report of achievement that is broad enough to cover all the important educational outcomes—subject achievement, character outcomes and social adjustment, health, and use of leisure.
5. Give an adequate picture of *causes* as well as of outcomes. . . .
6. Reflect a complete and sympathetic understanding of the child.
7. Afford a means of reporting flexible enough to account for the peculiar individual abilities of each child.
8. Give an account of pupil progress understandable and instructing to both pupil and parent.
9. Bring about closer coöperation and greater mutual understanding of home and school.
10. Provide for reciprocal reporting. [That is, space for suggestions and questions from the parent.]
11. Rate achievement in relation to the basic abilities and capacities of the child.
12. Rate achievement by means of valid and reliable marking systems.
13. Conform to reasonable standards of form and appearance. The report should be attractive.

The ordinary report card often fails to meet the fourth requirement in the above list. It tends to neglect the less tangible but important outcomes of education reflected in social and personal qualities. One advantage of the informal report card or letter to parents is that it attempts to inform parents on all phases of pupil growth. But it is the spirit of the report rather than its form which is important. Indeed a curt note from the teacher may be worse than the usual report card. Elsbree[24] cites the following letter from a teacher to the parents of a slow learner which is a good illustration of "How to Lose Friends and Influence Parents—in the Wrong Direction":

> Dear Parents,
>
> Donald has improved in nothing except spelling and that very little.
>
> Sincerely,
>
> ____________________
> Teacher

For use in the elementary school, Hill suggests the informal report to parents, reproduced with slight modifications in Figure 61. A similar form for the second half of the semester calls attention to improvements noted, and invites further parental co-operation on other points. Neither the report itself nor the letter accompanying it makes such demands upon the teacher's time as does the personal letter, which should probably be reserved for very special occasions.

---

[24] Willard S. Elsbree, *Pupil Progress in the Elementary School,* page 76. New York: Bureau of Publications, Teachers College, Columbia University, 1943.

REPORT FOR FIRST HALF OF THE FIRST SEMESTER

NAME -------------------- GRADE ------ ROOM------

PROGRESS IN SCHOOL SUBJECTS. ---------- is doing *very good* work in --------------------

His work is *good* in ----------------------------------

His work is *poor* and needs improvement in -----------

His work in these subjects would probably be improved if --

PHYSICAL CONDITION. Health habits and conditions needing attention -----------------------------------

ATTENDANCE. Half days absent ---- Number of times tardy ----

REMARKS. --

SCHOOL CITIZENSHIP. We believe that every boy should be happy in school, should take part in the life of the school, should get along well with his classmates, and should develop good habits of honesty, courtesy, neatness, consideration for the rights of others, and industry.

Your boy is especially strong in -----------------------

He could improve in ----------------------------------

- -

Tear off here and return this part of the sheet.

I have examined -------'s report for September and October.

Signed ----------------------

(Parent or guardian)

REMARKS OR SUGGESTIONS. -----------------------

Figure 61. A Suggested Informal Report to Parents. (After Hill.)

It is always a good idea, of course, to apply the grease *when* the squeak appears. The letter suggested to accompany the first report is as follows:

> Dear (*name of parent or guardian*):
>
> Now that the semester is one-half over we wish to call your attention to ....'s school progress. The enclosed report covers four kinds of progress—progress in school subjects, health and physical condition, attendance, and school citizenship. If you would like to talk over the report, or to get more complete information on your boy's success in school, we should be glad to have you come to see us. If you can telephone us or send a note ahead of time, it will make it easier to arrange a meeting.
>
> The upper part of the report is for you to keep for future reference. *Please return only the lower part.* We are especially anxious to get any information from you that will aid us in helping your boy make a complete success of his school work. Any information or suggestions you may wish to write will be welcome.
>
> Sincerely yours,
> (Signed by teacher and
> principal)

Suggestions for letters to parents. The art of writing effective letters to parents will require special training and practice. To assist teachers in acquiring this necessary skill, the schools of Santa Monica, California, prepared a very helpful list of suggestions.[25] The list in somewhat abridged form is as follows:

1. Begin the letter with encouraging news.
2. Close with an attitude of optimism.
3. Solicit the parents' cooperation in solving the problems, if any exist.
4. Speak of the child's growth—social, physical, and academic.
   a. Social (citizenship traits)
      (1) Desirable traits: attention. care of property, co-operation, honesty, effort, fair play, etc.
      (2) Undesirable traits: selfishness, wastefulness, untruthfulness, dishonesty, carelessness, etc.
   b. Physical (health conditions)
      Posture, weight, vitality, etc.
   c. Academic
      (1) Interest in school and extra-school activities.
      (2) Methods of work.
      (3) Achievements: (a) Growth in knowledge, appreciation, techniques; (b) list subjects in which child is making progress and those in which he is not making progress; (3) relationship of his accepted standards to his capacities.
5. Compare the child's efforts with his own previous efforts and not with those of others.
6. Speak of his achievements in terms of his ability to do school work.
7. Please remember that every letter is a professional diagnosis, and therefore is as sacred as any diagnosis ever made by any physician.

---

[25] *Ibid.*, pages 83–84.

A more elaborate 21-page manual to guide teachers in the preparation of reports to parents has been prepared by the Omaha, Nebraska, school system.

The Colorado experiment. Although it is true that the aim of all evaluation and reporting to parents is the complete development of the child, it is often necessary to "temporize ideals with practical considerations." [26]

The experience of the Secondary School of Colorado State College of Education is especially instructive.[27] The director reports that detailed analytical evaluation sheets were tried and abandoned primarily because of the excessive amount of time required to prepare them. The use of the terms *unsatisfactory, satisfactory,* and *honors* was given up because it was felt that any attempt to evaluate pupils both in terms of their own ability and the objectives of the curriculum is sure to involve negative reactions. Evaluations of the ordinary scale type were tried and abandoned because they afford only a partial report. Anecdotal records were attempted and discontinued because the teachers tended to select unusual activities and experiences instead of reporting an ordinary picture of the pupil's growth and progress. Conference meetings of counselor, teacher, and parents, although successful for a time, had to be given up because of the failure of the majority of parents to respond to the school's invitation to avail themselves of these conference opportunities. The school eventually prepared lists of "statements of trait actions" which were indicative of the pupil's attainment of such general school objectives as self-direction, social adjustment, breadth of interests, personal attractiveness, care of materials and equipment, basic reading skills, and the like. These were then evaluated on a five-point scale, *H,S,N,U,O,* indicating distinctly superior, satisfactory, needs to make improvement, unsatisfactory, and no evaluation, respectively.

The experiment has continued for thirteen years but the end is not yet. Recently Wrinkle [28] summarized the program as follows:

> In the thirteen years which have elapsed, new forms and new practices have been developed, tried, scrapped, and replaced by newer forms and practices. Detailed analytical reports, scale-type evaluations, the conference plan, anecdotal reports, and check-list type reports were developed and discarded because they did not do a good job of conveying information or demanded too much time.
>
> Repeatedly it was discovered that adequacy meant detail and detail meant

---

[26] See pages 89–90 in Chapter III.

[27] William L. Wrinkle, "The Story of a Secondary-School Experiment in Marking and Reporting," *Educational Administration and Supervision,* 23: 481–500, October, 1937.

[28] William L. Wrinkle, "Reporting Pupil Progress," *Educational Leadership,* 2: 293–295, April, 1945.

forms which were impractical for use in public school situations. One criterion which resulted in the scrapping of many forms and practices including those which were successful in their use in the laboratory school was: *Whatever is developed must be usable in the public schools by public school teachers.*

In May, 1945, a popular referendum was held in which all high-school students participated; the general consensus was highly favorable but several changes were proposed. For example, 99 per cent of the students thought they should always be allowed to see their scores on standardized achievement tests; also 90 per cent of the students thought that the reports to parents should show how the actual achievement compared to the expected achievement.

The University of Chicago High School system of reporting. The University of Chicago High School illustrates a dual system of reporting. At the end of each semester the parents receive a detailed report in terms of the specific objectives of each course and whatever comments are deemed necessary. A week or so after the detailed reports are sent out and the parents have had an opportunity to study the strengths and weaknesses of the pupil, the course marks are forwarded and are usually accepted by the parent as incidental supplementary information. Figure 62 illustrates one of the detailed semester reports in social studies. The Chicago system may be regarded as a desirable transition between the formal report cards and the informal letter to parents.

E. Other Avenues of Public Information

School exhibits. There is no sounder principle of evaluation than that contained in the statement, "By their fruits ye shall know them." Exhibits afford one of the best ways of presenting the "fruits" of the school. The public is evidently interested in local, county, state, national, and international fairs and exhibitions of all types, and schools could make use of this fact. Posters and displays of pupils' work, as well as public programs of a dramatic, literary, or musical character, afford concrete demonstrations of the school's educational program. Commencement programs in which the pupils themselves play the leading roles afford an excellent opportunity for the public to see the end products of the school. In the final analysis, however, the ordinary everyday behavior of the pupil is the best evidence of the worth of the school. What the pupil *thinks* and what the pupil *says* are both important; but what the pupil *is* speaks a still more eloquent language.

School visitation. Vicarious knowledge is important, but it is usually a poor substitute for first-hand experience. Whenever possible, therefore, the public should have an opportunity to see their school in actual operation. The school should cultivate a reputa-

THE UNIVERSITY OF CHICAGO

The Laboratory School

SEMESTER REPORT, SOCIAL STUDIES III----------

Student ------------------------------------- Date ---------------
Last Name First Name

Purposes	Rating	Comments (if any)
1. Acquisition of basic information		
2. Reading skills		
a. recognizing main ideas		
b. recognizing pertinent data		
c. social studies vocabulary		
3. Oral Skills		
a. presentation of ideas		
b. organization of ideas		
c. adequacy of content		
4. Writing Skills		
a. organization of ideas		
b. adequacy of content		
5. Ability to interpret social data		
6. Ability to apply principles in new situations		
7. Interest in current affairs		
8. Courtesy and cooperation in group situations		
Habits of Work		
9. Persistence in overcoming difficulties		
10. Tendency to work independently		
11. Promptness in completing work		
12. Application during study		
13. Attention to class activities		
14. Participation in class activities		
15. Effectiveness in following directions		

Pupil's Grade ---------- ----------------------
Instructor

Figure 62. A Report Card Used at the University of Chicago High School.

tion for friendliness. The announced policy of the school should be, "The latch string is always out." It is a rare parent indeed who would not rather see his own child "perform" than witness world-famous actors on the silver screen. Furthermore, to observe the process of upholstering a chair or fashioning a dress is inherently more interesting than merely to look at the finished product. One of the best ways of providing opportunities for observation is to hold "open-house" or "back-to-youth" programs in the evening, where parents can follow their children's daily activities, usually on a half-time schedule.

The parent-teacher association. The modern educator recognizes more clearly than did his predecessor that education is a continuous unified process, that several agencies contribute to its accomplishment, and that of these the home and the school are most important. It is self-evident, therefore, that there should be intelligent and wholehearted co-operation between the home and the school. The local parent-teacher association seeks through mutual understanding to effect this needed co-operation. At its best, the association is a modern successor to earlier visits of teachers to the pupils' homes and of the parents to the school, both of which are increasingly difficult with the growth of the school population and with the enlargement of the area served by the individual school.

From the viewpoint of the home, the association affords an opportunity for parents not only to *hear* about the school's program and philosophy and to *see* the school in actual operation, but also to *react* to what they hear and see. The modern parent, like the modern child, wants to be heard as well as seen. Certainly at all times he is entitled to a respectful hearing. When he is right, the school should make the needed adjustments; and when he is wrong, the school should attempt to set him right.

A successful school man suggests the following parent-teacher association program for the junior high school: [29]

I. Introduction of program by head of the mathematics department, who will briefly state
   1. General aims of mathematics offered in junior high school (grades seven to nine)
   2. Courses offered, as arithmetic, general mathematics, industrial arithmetic, and algebra
   3. Scope and aims of each course

II. Explanation of school's standing and accomplishment in mathematics—as shown by grade-level achievements as compared to national norms; display and explanation of exhibits; also achievement charts showing progress within the school over a given period of time—by a teacher

---

[29] L. N. Morrisett, *op. cit.*, pages 484–485.

III. Junior-high-school arithmetic from an eighth-grade pupil's viewpoint
    1. Aims and purposes
    2. How it functions
    3. What we learn
    4. Our project

IV. A twenty-five-minute lesson in eighth-grade mathematics
    1. This lesson from this day's program—an actual classroom recitation, developing and bringing out
        (a) Purposes of the recitation unit—with acceptance of same by class
        (b) Method—(socialized recitation)
        (c) Emphasis on the objectives
        (d) Drill
        (e) Relation and practical application
        (f) Summary

V. Round-table discussion; leaders: head of mathematics department and a patron
    1. An honest effort to relate the mathematics taught in school to the mathematics used in the life of the community
    2. What parents expect from the mathematics department
    3. Discussion of method used
    4. Discussion of invited questions

The program above is a truly co-operative affair, the outcome of which is sure to be improved public relations. It provides for the mutual exchange of ideas, which is the basis of true understanding. It aims at evaluation and not at mere information. To this end the measurement of results makes an important contribution.

F. Mobilizing Public Opinion

Sampling the opinion of parents. To what extent can the judgment of parents be utilized in the evaluation and improvement of the school? Eells [30] attempted to use the opinions of the parents of seniors in evaluating the secondary schools attended by their sons or daughters. He employed a five-point scale ranging from "extremely satisfactory" at one end to "extremely unsatisfactory" at the other. Twelve items were included, relating to the general quality of instruction, development of good character, training in good citizenship, guidance activities, and the like. The principal of the school personally signed and mailed to the parents of seniors in his school a return postcard containing the following message:

> *To the Parents of Seniors:*
> Our school has been selected as one of two hundred high schools and other secondary schools in the United States to be critically studied and evaluated in an effort to improve the standards of secondary education throughout the country. The study is not connected in any way with the Federal Government.

---

[30] Walter Crosby Eells, "Judgments of Parents Concerning American Secondary Schools," *School and Society*, 46: 409–416, September 25, 1937.

> One part of the plan for this national study calls for a frank evaluation of the school from the standpoint of the parents. We are asking parents of our seniors to state their honest opinions concerning certain aspects of our school, as judged by the development of their children during their school life here. You are urged to express your candid judgment, whether it is favorable or unfavorable. You are not asked either to praise or to defend the school, only to judge it. The card need not be signed and it is to be sent directly to the headquarters of the study in Washington. I shall not see it again.
>
> I am eager to have a hundred per cent response from the parents of pupils in this school. Won't you fill the card out and mail it promptly? Within a day or two, please!

The study concluded that "the parents, on the whole, showed a marked degree of discrimination," as judged by the scattering of the ratings along the scale. Only a quarter of the ratings were "exceedingly satisfactory," and more than 7 per cent were "not very satisfactory" or "exceedingly unsatisfactory." It is significant that the guidance program of the typical school was considered least satisfactory, a judgment supported by other criteria. Yet perhaps no phase of the school program is more dependent for its success upon parental co-operation than is guidance. "Regardless of whether parents are correct in their judgments, it is important to know what these judgments are," Eells points out, "for in the last analysis the parents support and control the schools." Another writer[31] emphasizes the point that although it is important to discover what the public *knows* about its schools it is even more important to learn what the public *feels* about its schools.

Concluding statement. It is one of the fundamental beliefs of a democracy that reliance can be placed on an enlightened public opinion. It is to achieve this end that public schools are maintained. But it is erroneous to assume that the responsibility ceases when the formal period of instruction ends. In a changing world the continued enlightenment of the adult population is increasingly recognized as a major responsibility of a democratic society. No individual or group can be expected to think or to act intelligently on anything without the necessary information. To supply this information about the schools is the objective of the public relations program. At all times the school will do well to keep in mind the words of one of America's ablest statesmen, Abraham Lincoln:

> Public sentiment is everything. With public sentiment nothing can fail, without it nothing can succeed. Consequently he who molds public sentiment goes deeper than he who enacts statutes or pronounces decisions.

---

[31] Warren C. Seyfert, "What the Public Thinks of Its Schools," *School Review,* 48: 417–427, June, 1940.

SELECTED REFERENCES FOR FURTHER READING

Elsbree, Willard S., *Pupil Progress in the Elementary School.* New York: Bureau of Publications, Teachers College, Columbia University, 1943. Chapter VIII.

Evans, Robert O., *Practices, Trends, and Issues in Reporting to Parents on the Welfare of the Child in School.* New York: Bureau of Publications, Teachers College, Columbia University, 1938. 98 pages.

Farley, Belmont Mercer, *What to Tell the People about the Public Schools.* New York: Bureau of Publications, Teachers College, Columbia University, 1929. 136 pages.

Grinnell, J. Erle, *Interpreting the Public Schools.* New York: McGraw-Hill Book Company, Inc., 1937. 360 pages.

Moehlman, Arthur B., *Social Interpretation.* New York: D. Appleton-Century Company, 1938. 485 pages.

Reeder, Ward G., *An Introduction to Public-School Relations.* New York: The Macmillan Company, 1937. 260 pages.

Smith, Eugene R., Tyler, Ralph W., and Staff, *Appraising and Recording Student Progress,* New York: Harper & Brothers, 1942. Chapters IX-XI.

Traxler, Arthur E., *Techniques of Guidance.* New York: Harper & Brothers, 1945. Chapter XIII.

Woody, Clifford, and Sangren, Paul V., *Administration of the Testing Program.* Yonkers: World Book Company, 1933. Chapter IX.

Wrinkle, William L., and Gilchrist, Robert S., *Secondary Education for American Democracy.* New York: Farrar and Rhinehart, 1942. Chapter 39.

PUBLISHERS OF STANDARD TESTS

American Council on Education, 744 Jackson Place N. W., Washington, D. C.
Association Press, 347 Madison Avenue, New York City
Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas
Bureau of Educational Research and Service, State University of Iowa, Iowa City, Iowa
Bureau of Publications, Teachers College, Columbia University, New York City
California Test Bureau, 5916 Hollywood Boulevard, Los Angeles, California
Center for Psychological Service, George Washington University, Washington, D. C.
Character Research Institute, Washington University, St. Louis, Missouri
C. H. Stoelting Company, 424 North Homan Avenue, Chicago, Illinois
Committee on Publications, Harvard Graduate School of Education, Cambridge, Massachusetts
Co-operative Test Service, 15 Amsterdam Avenue, New York City
Division of Educational Reference, Purdue University, Lafayette, Indiana
Educational Test Bureau, 720 Washington Avenue, S. E., Minneapolis, Minnesota
Extension Department of the Training School, Vineland, New Jersey
C. A. Gregory Company, 345 Calhoun Street, Cincinnati, Ohio
Harvard University Press, Cambridge, Massachusetts
Houghton Mifflin Company, 2 Park Street, Boston, Massachusetts
Keystone View Company, Meadville, Pennsylvania
Ohio College Association, Ohio State University, Columbus, Ohio
Psychological Corporation, 522 Fifth Avenue, New York City

Psychological Institute, Washington, D. C.
Public School Publishing Company, 509–513 North East Street, Bloomington, Illinois
Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois
Scott, Foresman and Company, 623 South Wabash Avenue, Chicago, Illinois
South-Western Publishing Company, 201–203 West Fourth Street, Cincinnati, Ohio
Stanford University Press, Stanford University, California
University of Chicago Press, Chicago, Illinois
West Publishing Company, St. Paul, Minnesota
World Book Company, Yonkers, New York

INDEX

Index

A

Abbott, Allan, 147
Ability, meaning of, 285
Ability grouping, 437–445
  arguments for and against, 438
  experiments, 438–441
  technique, 441–443
Abnormal psychology, France and, 31–36
Acceleration and retardation, 443–445
Accomplishment quotient, 309–310
Accuracy of measurement, 66
Achievement:
  intelligence and:
    combining, 310
    comparing, 308–310
  levels, and intelligence group, 458–459
  tests (*see* Achievement tests)
Achievement tests, 24
  after 1918, 49–52
  Armed Forces Institute, 283
  battery of, locating retarded readers, 387
  before 1918, 42
  first, 44
  general, class record, 220
  history of, 42–52
  local norms, value of, 307
  movement, 21–22
  nonstandardized, advantages and limitations, 184–185
  objective, 50
    origin of, 51
  progressive, profile of student on, 261
  reports to pupils, 207–208
  scores, combining with intelligence scores, 308–310
  standardized:
    advantages and limitations, 184–185
    individual profile chart, 210
  studies of school marks and examinations, 44–49
  types of, after 1927, 50
  University of Minnesota, 105
  use of norms in interpreting scores, 301–307
  validity of, 70–82 (*see also* Validity: achievement tests)
  when to administer, 195
Ackerson, Luton, 86
Activity movement, 435–437
Adams, A. Elwood, 412
Adams, Eunice, 305
Adams, Jesse E., 443
Adjustment, 3–4
Administering tests, 193–199
  ease of, 90–91
  procedure, 197–199
  time for, 193–195
  who shall administer, 195–196
Administration, school, function of measurement in, 22–24
Administrators, records and reports for, 209–214
After effects of specific connections, 339–341
Age:
  chronological, 237, 242, 287 *ff.*
  educational (*see* Educational age)
  increase, 67–68
  mental (*see* Mental age)
Alden, Clara L., 387
Alienation, coefficient of, 243
Allen, Mildred M., 194
Allen, Richard D., 450, 455, 471, 484
Allport, Gordon W., 56
Alpha scale, 491, 492, 501, 506
Alternative-response tests, 127, 139–144
  advantages and limitations, 139–140
  construction, 143–144
  defined, 139
  illustrations, 141–143
Altstetter, M. L., 449
Amateurs, and new-type test, 104
America, and applied psychology, 36–37
American Council Psychological Examination, 474
American Journal of Psychology, The, 58
American Psychological Association, 40
Analogy tests, 127
Analysis:
  qualitative to quantitative, trend from, 8
  statistical, of test results, 216–251 (*see also* Statistical analysis of test results)
Anastasi, Anne, 85
Anderson, C. J., 318
Anderson, Harold A., 160
Anderson, H. Dewey, 466
Anderson, H. R., 93
Anderson, V. V., 465
Andreen, Earl P., 360
Annual reports, 519–520
Answer keys, 123
Application of tests, ease of, 92–93
Applied psychology, America and, 36
Applied sciences, 5
  measurement in, 10–12
Appraisal (*see also* Evaluation)
  guidance, 451
Approximation, 12
Aptitude tests, 41, 61
  and predicting achievement, 479
  O'Rourke on, 70
AQ (accomplishment quotient), 309–310
Areas of investigation, 375–376
Aristotle, 7, 9, 14, 21
Arithmetic:
  mean, 230
  paper, study in grading, 47
Armed Forces Institute, achievement tests, 283
Army Alpha tests, 17, 40, 42, 67
  scores, distribution of, for 5 occupations, 460
Army Beta test, 40, 43

Army General Classification Test, 41, 425
Arnold, Dwight L., 108
Art:
  creative expression in, 109–111
  science and philosophy, 16
Ascendance-submission, measurement of, 56
Ashbaugh, Ernest J., 47, 85
Ashburn, Robert R., 158
Asher, E. J., 286
Astronomy, 6
Atomic physics, 17
Attitudes, measurement of, 56, 57
Authorities, 20–21
Automatic scoring, 356–358
Automobile, 11
Avent, Jos. E., 104
Average:
  mean, 230–233
  measures of, 227–233
  median, 228–230
  mode, 227–228
  what is best, 233
Ayres, Leonard P., 43, 44, 51
Ayres educational index, 76

B

Back-to-youth programs, 529
Baker, Arthur O., 133
Baker, Harry J., 62, 289, 366, 369, 383
Bamberger, Sister Clara Francis, 258
Bar graphs, 255–257, 259, 270
Barnes, Harry Elmer, 10
Barr, A. S., 22, 90, 493
Barrett, E. R., 147
Barton, W. A., Jr., 139
Batteries of tests, 87, 184
  achievement tests, 387
Bayless, Ernest E., 139
Bear, Robert M., 389
Beauchamp, W. L., 133, 136, 142
Beck, Roland L., 135
Bedell, Ralph C., 139
Beggs, V. L., 522
Behaviorism, 17
Benjamin, Harold, 449
Beta scale, 492, 501
Betts, Emmett Albert, 387, 393
Bible, 27
Bibliographies, 61
  of available tests, 370
Bibliography of Mental Tests and Rating Scales, 61
Billett, Roy O., 397, 399, 401, 404, 408, 409, 424, 429, 438, 442
Binet, Alfred, 30, 31, 33–36, 59
Binet scale, 34, 36–37, 40
Bingham, Walter V., 58, 425, 451, 461, 462, 469, 470, 481
Biological sciences, 5
  measurement in, 7–9
Bixler, Harold H., 397
Bixler, Ray H., 456
Bixler, Virginia H., 456
Black, Max, 14
Blair, Glenn Myers, 360, 367, 370, 393
Blin, 34
Bobbitt, Joseph M., 378
Book, William F., 315, 335, 413
Books, important, 59–61
Bordin, Edward, 109
Boring, Edwin G., 6, 15, 34
Boss, Mabel E., 521
Boston, first written educational report, 519
Botany, 6
Boyer, Philip A., 179, 484, 521
Boynton, Paul L., 198, 289, 290, 296
Brenner, Benjamin, 344
Brinkmeier, I. H., 114
Brinton, W. C., 280
Brown, Clara M., 148, 466
Brown, Francis J., 331
Brown, F. W., 152
Brown, Marion, 481
Brownell, William A., 14, 86, 131, 192, 364, 378
Brubacher, John S., 15
Brueckner, Leo J., 89, 371, 373
Bruner, Herbert B., 502–504
B-scores, 305
Buckingham, R. B., 15, 49, 50
Burbank, Luther, 311
Buros, Oscar Krisen, 61, 81–82, 186, 188
Burt, Cyril, 31
Byrns, Ruth, 478

C

CA (Chronological age), 237, 242, 287 *ff.*
Cady, 55
Caldwell, Otis, 29, 43
Caldwell, V. V., 430
California, survey of marking, 398
Camden, N. J., schools, costs, 253–254
Cameron, Dale C., 378
Camp, 322
Campbell, Pera, 378
Cantril, Hadley, 58
Capacity, meaning of, 285
Carr, Harvey, 349
Carter, Harold D., 78
Carter, Ralph E., 163, 164
Case studies, diagnosis in reading, 383–386
Cason, Hulsey, 341
Cattell, James McKeen, 30, 36
  "Mental Tests and Measurement," 58
Cattell, Psyche, 294, 298
Centile score, 299–300
Central tendency:
  allowing for variations in, marks, 417–421
  graphs representing, 274–278
  measures of, analysis, 227–233
Chadwick, E., 515
Chalmers, T. M., 328
Character and personality measurement, 24
  beginnings, 52–53
  history of, 52–58
  later development, 53–55
Character education, 53
Character Education Inquiry, 53
Charts, diagnostic, 380–381
Cheating, 356–359
Check lists, 25
Chemistry, 6
Chicago study of high-school grades, 45
Childers, Leon M., 47
China, examination system of, 27–28
Chronological age (CA), 237, 242, 287 *ff.*
Circle graphs, 258, 270
Clapp-Young self-marking tests, 91
Clark, Champ, 19
Clark, E. L., 424
Clark, Harold F., 467

Classification, 7–8
  and promotion, 423–446
  differentiated unit assignments (*see* Differentiated unit assignments)
  group differences, 424–426
  human variability, 423–430 (*see also* Human variability)
  individual differences, 426–428
  provisions for, 429–430
  test results, 218–227
    rank order, 219–222
  trait variability, 428–429
Class records:
  general achievement test, 220
  reading readiness test, 219
Class tests, 351–360
  devices for increasing value of, 355
  kinds of, 351–352
Client-centered counseling, 450–451
Cobb, Irvin S., 19
Coble, Robert, 400
Cochran, Roy E., 159, 168
Coefficient:
  of alienation, 243
  of correlation, 237–246
    computing methods, 238–241
    interpreting, 241–244
    uses of, 244–246
  validity, 244
Coherency, 67, 68
Cole, Robert D., 189
Cole-von Borgersrode scale for rating standardized tests, 189–192
College achievement, predicting, 474, 475–481
College Board Examinations, 318
Colleges:
  evaluation, 497–498
  marking, before 1918, 45
  tests in, 23
Column diagram, 265, 266
Common sense, 68
Community evaluation, 498, 499–500
Comparative examinations, 52
Completion tests, 37, 127
  advantages and limitations, 135
  construction, 136–138
  defined, 134
  illustrations, 135–136
Composition, English, estimated by teachters, *table*, 46
Compton, B. K., 335
Concentration, 110
Conklin, Edmund S., 428
Conneau, 127
Conner, 323
Conrad, H. S., 308
Consistency, 83
Constant, personal, 298–299
Construction:
  essay examination, 162–167
  graphs, 278–281
  tests:
    achievement, 75–79
    alternative-response, 143–144
    completion, 136–138
    general principles, 103–126
    importance of problem, 103–104
    objective, specific types of, 127–156
    teacher-made, 101–171
Controlled observation, 58
Control over universe, 4
Cook, Walter W., 87, 119, 130, 515
Cooking, 11–12
Co-operation, 111, 339
Cooperative Achievement Tests, 72, 105, 303
Cooperative Study of Secondary School Standards, 489, 490, 491, 493–495, 498 *ff.*
  instructions for using criteria, 501
Co-operative testing program, 179
Cornell, Ethel L., 439
Cornell, Francis G., 509–513
Cornog, J., 133
Correction formula, 119, 120, 122
Correlation, 245–246
  coefficient of, 237–246 (*see also* Coefficient: of correlation)
  product-moment of, Pearson, 69
  technique, 31, 36
Correlational psychology, 59
Costs:
  current, in states, *graph*, 256–257
  per capita, Camden schools, *graph*, 253–254
  testing, 93–94
Counseling, 450, 452, 456–462
  educational guidance, 479–481
  information needed for, 457
  knowledge needed by teacher, 455–456, 457–458
  nature of, 456
  vocational guidance, 470–471
Courtis, Stuart A., 29, 43, 68
Courtis practice tests, 360
Cowdry, 56
Cox, Philip W. L., 471
Cramming, 318, 327
Crane, Margaret, 449
Crawford, Albert Beecher, 410
Crawford, C. C., 104
Crawford, John R., 302
Creativeness, 109–111
Cressey, Paul F., 27
Critical ratio, 250
Critical thinking, 108–109
Cronbach, Lee J., 119
Crossley, Elizabeth, 71–72
Crude scores, 286
C-score, 300, 306–307
Cuff, Noel B., 92
Cunliffe, Rex B., 468, 470, 471
Cureton, Edward E., 309
Curricular validity, 70
  achievement tests, 70–71
Curtis, Francis D., 139, 355
Curves:
  in graphs, 267–270
  normal, 269–270
  percentile, 268–269
  skewed, 269–270
  smooth, 267–269
Cutright, Prudence, 353

D

Dalton plan, 430, 432–434
Daly, Joseph F., 246
Dampier-Wheham, William C. D., 68
Darley, John G., 402, 450, 462, 465, 470, 482
Darling, W. C., 139
Darwin, Charles, 7–8
Data, testing ability to use, 108–109
Davidson, Percy E., 466
Davis, Georgia, 378
  spelling, difficulties and remedies, 379
Davis, J. De Witt, 411
Davis, Robert A., 424

Davison, F. M., 419
Dearborn, Walter F., 89, 200, 290
Dentistry, examination papers, regraded, *table*, 48
Deputy, E. C., 331–332, 333
Derived scores, 285–287
  defined, 287
Detroit Advanced Intelligence Test, 69
Deviation of scores:
  quartile, 234, 235
  standard, 234–237
Dewey, John, 4, 11, 436
Diagnosing Personality and Conduct, 54
Diagnosis, 364–396
  areas of investigation, 375–376
  difficulties, locating, 370–374
  errors:
    analysis of, 370–374
    causes of, locating, 374–378
  individuals needing, 368–370
  in reading, 382–395
    case studies, 383–386
    techniques, 386–395
  intelligence tests, role of, 366
  levels of, 368
  nature of, 364–366
  preventive, 382
  problem of, 364–368
  techniques, 368–382
  value of, 366–368
Diagnostic tests, 50
Diamond, Leon N., 80
Dickey, John W., 217–218
  study of tests, 216–218
Dictionary of Occupational Titles, 465
Diederich, Paul B., 209
Differences:
  group, 424–426
  individual, 426–428
Differentiated unit assignments, 430–437
  activity movement, 435–437
  Dalton plan, 432–434
  Morrison plan, 435
  Winnetka plan, 430–432
Direct vs. indirect methods, achievement tests, 76–78
Distributions, series of, graphs, 272–278 (*see also* Series of distributions)
Dolbear, Amos E., 429
Doughton, Isaac, 16
Douglass, Harl R., 342
Drill machines, 356–358
Drives, 316
Duff, John Carr, 471
Dunlap, Knight, 41, 349
Durant, Will, 16
Durrell, Donald D., 371, 387
Dykema, Leon N., 80

E

EA (*see* Educational age)
Eaton, Merrill B., 21
Ebbinghaus, Hermann, 31, 37
Ecological studies, 8
Economy Remedial Exercise Cards, 360
Edgerton, A. H., 463, 468
Edmiston, R. W., 166
Education:
  measurement in, 14–26
    conflicting views of, 17–19
    historical development of, 27–64
    place of, 16–19
  nature of, 14–16, 447
  reputation and, 19
Education (*Cont.*):
  research and, 19
  rhetoric and, 19
  three R's in, 19–22
  types of measurement in, 24–26
Educational, Psychological, and Personality Tests of 1933, 1934, and 1935, 61
Educational Administration and Supervision, 59
Educational age (EA), 226–227, 237, 242
  limitations of, 302–304
  uses of, 302
  vs. educational quotient, 301–302
Educational diagnosis (*see* Diagnosis)
Educational guidance, 53, 472–482
  counseling, 479–481
  opportunities and requirements, analysis of, 473–475
  placement, 481–482
  vocational vs., 472–473
Educational program, evaluation of, 499–513
Educational quotient:
  educational age vs., 301–302
  limitations of, 304
  use of, 304
Educational Test Bureau, 262
Edwards, I. Newton, 423
Eells, Walter C., 449
Effect, principle of, 347–348
Efficiency:
  of school system, 511
  of teachers, evaluating, 491–493
Eight-year study, Progressive Education Association, 78, 105, 489–490
Einstein, Albert, 7, 315
Elementary schools, evaluation for, 495–496
Elliott, Edward C., 45
Ellis, Albert, 484
Ellwood, Charles A., 10
Elsbree, Willard S., 523
Elwell, Mary, 89, 371
Emphasis, teaching, and measurement, 317–319
Engelhart, Max D., 249
Engelhardt, N. L. Jr., 513
England and statistical methods, 31
Engle, Thelburn L., 444
English composition:
  estimated grade-value and percentage marks, *table*, 46
  part of Hudelson Scale, 47
Erickson, Clifford E., 264
Errors, 12–14
  analysis of, 370–374
  causes of, locating, 374–378
  controlling, methods of, 13
  in diagnosis, 370–378
  in scoring tests, 200–201
  interpretation of, 249–250
  IQ subject to, 298
  of measurement, 95–97
  types of:
    measurement, 247–248
    sampling, 248–249
    technique, 246–247
Essay examination, 157–171
  advantages of, 160–162
  improving, 162–171
    construction and use, 162–167
    scoring or grading, 167–171
  limitations of, 157–160

Essay examination (*Cont.*):
  marking, 168–169
  types of, 162–163
Eurich, Alvin C., 105, 129, 188, 325, 491
Evaluation, 19, 488–514
  community, 498, 499–500
  Cooperative Study, 493–495
  difficulty of, 491
  educational program, 499–510
  elementary schools, 495–496
  higher institutions, 497–498
  importance of, 490–491
  measurement and, 488–490
  philosophy of school, 498–499
  principles of, 495–498
  problem of, 488–495
  school organization and plant, 510–513
  secondary schools, 496–497
    scales for, 492
  teacher efficiency, 491–493
  teacher-made tests, 123–125
    checking of, 124–125
    reliability coefficient, 125
    validity, 124
Evaluative Criteria, 498–499, 501, 503, 504, 506, 507, 510, 511
Evans, Robert O., 522
Evolution, 7–8
Examinations:
  comparative, 52
  final:
    awareness of, 326–329
    value of, 361–363
  improved, 51
  of Chinese, 27–28
  papers, dentistry, regraded, *table*, 48
  studies in unreliability of, 44–49
Examiner, personality of, 96–97
Exercise, principle of, 346–347
Exhibits, school, 527
Experimental psychology:
  Germany and, 30–31
  publications, 58–61
Experimental research, 22
Experimental science, beginnings of, 4, 6
Extroverted pupils, 337–338

F

Fabre, Jean Henri, 8
Fairmindedness, Watson's measurement of, 56
Falls, J. D., 46
Farley, Belmont Mercer, 517
Fatigue, 37
Fechner, 31
Feder, Daniel, 484
Fei Tsao, 309
Ferguson, G. A., 83
Fernald, G. G., 53
Fernald, Grace M., 382
Final examination:
  awareness of, 326–329
  value of, 361–363
Finch, F. H., 78
Fish, Louis J., 520
Fisher, Reverend George, 43–44
Flanagan, John C., 298
Flesher, Marie A., 444
Folk, S. B., 444
Follow-up:
  educational guidance, 481–482
  guidance, 462–463
  vocational guidance, 471–472
Forlano, George, 330, 340, 344, 354
Formal tests, 24
Form board, 40
Fowler, Fred M., 451–452
France, and abnormal psychology, 31–36
Frandsen, Alden, 469
Franzen, Raymond, 309
Freeburne, Max, 354
Freeman, Frank N., 299, 360
Freeman, Frank S., 428
Frequency distribution:
  column diagram, 265
  curves, 267–270 (*see also* Curves)
  graphs, 264–271
  histogram, 265
  polygon, 265
  smooth curve, 267–269
Frequency of tests and motivation, 322–326
Frequency polygon, 265–267
Frequency table, 222–227
  making, 222–225
  scattergram, 225–227
  two-way, 225–227
Fryer, Douglas, 457, 469
F-score, 309
Furfey, Paul Hanly, 246

G

Gable, Sister Felicita, 323
Galen, 53–54
Galileo, 4, 6, 11
Gallup, George H., 58
Galton, Sir Francis, 8, 30, 31, 33, 36, 44, 58
  on character measurement, 52–53
  questionnaire, 55
Gamma scale, 492, 501
Gard, Paul D., 91
Garrell, H. E., 301
Gates, Arthur I., 147, 295, 354, 437
General intelligence testing, 30–41
Geology, 6
Geometric Aptitude, Lee Test of, 459–461
Germany, and experimental psychology, 30–31
Gestalt school of psychology, 8–9, 17, 18
Gillespie, F. H., 56, 57
Gilliland, A. R., 424
Gillis, Ezra L., 404
Givens, Meredith B., 465
Goals of attainment, 282–284, 316
Goddard, Henry H., 36
Goldenweiser, Alexander, 9–10
Gooch, Wilbur I., 465
Good, Carter V., 15, 22, 90, 474
Goodenough, Florence L., 86
Gordon, Hans C., 521
Grade norms, 304–306
Grading:
  essay examination, 167–171
  of students, 44–49
Graduates and nongraduates, high school, 271
Graphs, 252–281
  bar, 255–257, 259, 270
  circle, 270
  constructing, 278–281
  frequency distribution, 264–271
  pictorial, 254, 258, 270
  pie, 270
  profiles, 259–264 (*see also* Profiles)
  psychographs, 259
  record of individual, 259–264

Graphs (*Cont.*):
  series of distributions, 272–278 (*see also* Series of distributions, graphs)
  value of, 252–259
  which are best, 270–271
Gray, J. Stanley, 14
Gray, William S., 394
Greenberg, Jacob, 148, 154
Greene, H. A., 142
Gregory, C. A., 136
Grimes, James W., 109
Group differences, 424–426
Grouped frequency distribution, 222
Group instruction, and individual instruction, 437
Group tests, individual vs., 68–70
G-scores, 305
Guessing answers to tests, 119–120
Guidance, 18, 53, 366, 447–487
  adjustment, 451
  analysis of individual, 454–456
  appraisal, 451
  counseling (*see* Counseling)
  educational, 472–482 (*see also* Educational guidance)
  importance of, 447–450
  meaning of, 447
  opportunities available, analysis of, 453
  personal, 482–486 (*see also* Personal guidance)
  placement, 462
  place of measurement in, 450–452
  scope of, 451
  technique, 452–463
  use of profiles in, 262–264
  vocational (*see* Vocational guidance)
Guiler, Walter Scribner, 367
Guilford, J. P., 85

H

Hadley, Loren S., 476
Haggerty, Lida Harmar, 309
Haggerty, M. E., 497
Hall, G. Stanley, 55, 56, 520
  magazines founded by, 58, 59
Hall, Wilbur, 311
Handwriting, difficulties and remedies, 380–381
Happ, Marian Crossley, 264
Harap, Henry, 424
Harris, Albert J., 306
Harris, Daniel, 405
Harry, David P., Jr., 132, 155
Hartley, Henry H., 521
Hartshorne, Hugh, 53, 55
Hartson, L. D., 474
Hart test of social attitudes, 56
Harvard Committee, Report of, 490
Hawkes, Herbert E., 115, 121, 129, 131, 145, 150, 151, 162
Hawkinson, Mabel J., 373
Healy, William, 40
Heilman, J. D., 352
Heinis Mental Growth Units, 298, 300
Helmholtz, 8, 31
Henmon, V. A. C., 478
Henry, Lyle K., 324, 325
Heredity, 8
Herrick, James B., 12
Hertz, 11
Hertzberg, O. E., 352
Hevner, Kate, 139
Higher institutions, evaluation, 497–498
High school:
  measuring attitudes toward, 56, 57
  study of grades, 45
Hildreth, Gertrude, 61, 63, 298, 369, 373, 375, 387, 391
Hilkert, Robert N., 214
Hill, George E., 140, 522
Hill's study of report cards, 522–525
Hirshstein, Bertha, 78
Histogram, 265, 266
Historical research, 22
History:
  of achievement tests, 42–52
  of educational measurement, 27–64
Hoff, Arthur G., 358
Hoglan, 322
Hollingworth, H. L., 454
Homogeneous groups, 437–445
Horst, Paul, 150
Hubbard, Henry D., 252
Hudelson Scale, 47
Huey, 36, 59
Hull, Clark L., 60, 427, 428
Hulten, C. E., 47
Human variability:
  group differences, 424–426
  individual differences, 426–428
    provisions for, 429–430
  nature and significance of, 423–430
  problem of, 423–424
Hunnicutt, Clarence W., 337
Hurlin, Ralph G., 465
Hurlock, Elizabeth B., 336–337
Hu Shih, 4
Huxley, Julian, 7
Hyde, M. F., 440

I

Identification tests, 127
Idiot, 40
"If," 402–403
Imbecile, 40
Imus, Henry A., 389
Incentives, 316
Incorrect-statement tests, 127
Index:
  of studiousness, 308
  of variation, evaluation, 513–514
Indiana accrediting association, 76
Individual, analysis of:
  educational guidance, 475–481
  guidance, 454–456
  vocational guidance, 467–469
Individual differences, 30, 31
  educational provisions for, 429–430
  greater provisions for, *tables*, 431–434
Individual instruction, and group instruction, 437
Individual profile chart, Metropolitan Achievement Tests, 210
Individual tests, vs. group tests, 68–70
Informal tests, 24
Initiative, 109
Innate intelligence, 67
Instruction:
  function of measurement in, 22–24
  on college level, 23
Intelligence:
  and achievement, comparing, 308–310
  meaning of, 67
  Terman criteria, 67–68
  two-factor theory of, 59

Intelligence quotient (IQ), 18, 63
  advantages of, 294–295
  background of, 31
  computation of, 290–292
  equating, 296–297
  interpretation of, 292–294
  limitations of, 295–298
  mental age vs., 287–288
Intelligence tests, 24
  American contribution, 36–37
  correlation of group tests with criteria of validity, 69–70
  English contribution, 31
  French contribution, 31–36
  general intelligence, 30–41
  German contribution, 30–31
  group vs. individual, 68–70
  history of, 30–41
  pupils' knowledge of, 334–336
  recent trends, 61–64
  reports to pupils, 207–208
  scores:
    combining with achievement scores, 310
    use of norms in interpreting, 287–301
  specific intelligence, 41
  validity of, 66–70
  when to administer, 193–195
  yearly, 186
Interest, 110
  inventories, 469
  measuring, 55
Intermediate tests, 351–360
International Institute, Teachers College, 158
Interpretation of tests, 92–93
Interviews:
  personality measurement, 54, 56–58
  teacher-pupil, 377–378
Introversion-extroversion, measurement of, 56
Introverted pupils, 337–338
Iowa, high-school achievement in, 77
Iowa Academic Testing Program, 317
Iowa High School Contest Examination, 478
Iowa Placement Examinations, 41
Iowa Silent Reading Test, 188, 478
  frequency polygons, 272–273
  total comprehension scores, 274
IQ (*see* Intelligence quotient)
Irwin, J. O., 62
Irwin, M. E., 316
Ivins, Lester S., 515

J

Jackson, Jesse D., 86
Jackson, R. W. B., 83
Jefferson, Thomas, 423
Jenkins, William Leroy, 416
Jersild, Arthur T., 350
Jessup, Walter A., 77
Jesus, 423
Jevons, W. Stanley, 30, 31
Johnson, Bess E., 325
Johnson, Franklin W., 45
Jones, Edward Safford, 165
Jones, Harold E., 324
Jordan, R. C., 85
Journal of Educational Psychology, The, 59
Journal of Educational Research, 59
Journals, professional, 58–59
Judd, Charles H., 448
Judgment, 110–111

K

Kandel, I. L., 475
Karsten, Karl G., 281
Kaulfers, Walter Vincent, 473
Kay, Marjorie E., 374
Kefauver, Grayson N., 297, 447
Kelley, Truman Lee, 59, 60, 61, 86, 95, 124, 135 198, 308
Kelvin, Lord, 6
Kentucky:
  intelligence tests, 9th grade, 69–70
  mountain children, 286
  Shelbyville, report of public schools, 276, 279
  trends in enrollment, *graph*, 254, 258
Keys, Noel, 23, 324, 351, 366, 485
Kilpatrick, William H., 17, 18, 436
King, Ronald, 6
Kinney, L. B., 129
Kirkpatrick, James Earl, 323, 350
Kitch, Loran V., 323
Kitson, Harry D., 449, 453, 469
Kittle, Marian A., 218
Knowledge:
  of final examination, 326–329
  of test scores, 329–339
  subjective and objective, 334
Koerth, Wilhelmine, 315
Koos, Leonard V., 447
Kopel, David, 388, 391, 393
Kraeplin, Emil, 31
Kuder, G. F., 86, 469
Kugle, 323
Kuhlmann, Frederick, 36, 298
Kuhlmann-Anderson tests, 69, 194, 298
Kulp, Daniel H., II, 23, 325

L

Laird, 55
Lake View High School, Chicago, 264
Lamson, Edna E., 335, 398
Langmuir, Irving, 96
Language of tests, 114–116, 118–119
L' Année Psychologique, 33, 59
Larson, Agnes A., 466
Lawler, Eugene S., 258
Learned, William, 425
Learning:
  after effects of specific connections, 339–341
  amount and quality of, relation of measurement to, 322–341
  awareness of final examination, 326–329
  experiment in motivation, 319–322
  frequency of tests, 322–326
  knowledge of test scores, 329–339
  nature of, 346, 447
  negative, danger of, 349–350
  positive, evidence of, 350–351
  principles of, some important, 346–348
  relation of measurement to motivation in, 316–317
  type of, relation of measurement to, 341–343
Leary, B. E., 105
Lee, J. Murray, 103, 127, 129, 145, 162, 180, 181, 199, 456, 459–461, 482
Lee, Richard E., 4
Lee Test of Geometric Aptitude, 459–461
Lehman, 56

Leker, Charles A., 414
Leland, Bernice, 383
Lentz, Theodore F., Jr., 53, 78, 119
Leonard, E. A., 209
Leonard, J. Paul, 367
Leonard, Sterling A., 136
Letters to parents, 521, 525–527
  Colorado experiment, 526–527
  suggestions for, 525–526
  University of Chicago High School, 527, 528
Leunberger, H. W., 352
Life histories, 8
Life insurance companies, 81
Ligon, Ernest M., 197, 199
Lincoln, Abraham, 490
Linder, Ivan H., 398
Lindquist, E. F., 87, 93, 107, 115, 121, 129, 131, 145, 150, 151, 162, 249, 307, 363
Line graph, 259
Link, Henry C., 472
Literature, testing appreciation of, 107–108
Little, James Kenneth, 356
Locy, William A., 7, 8
Longstaff, H. P., 325
Loofbourow, G. C., 485
Lorge, Irving, 92
Lowell, A. Lawrence, 362
Lundholm, H. T., 132, 143

M

MA (*see* Mental age)
McCall, William A., 17, 23, 50, 59, 60, 196, 412, 413, 441, 442, 492
McCallister, James M., 387, 388, 389
McCall Multi-Mental test, 69
McClusky, Howard Yale, 361
McConn, Max, 362
  on the "new" tests, 28
McCullough, Constance M., 188, 388, 392
Machines, testing and drill, 356–358
McIntosh, John Ranton, 350
McKinney, H. T., 444
McNamara, Walter J., 123, 146, 150
McNemar, Quinn, 34, 58, 67, 294
Macrae, Angus, 465
Madsen, I. N., 200
Maller, Julius Bernard, 339, 440, 479
Mallett, Donald R., 484
Malthus, Thomas R., 10
Manipulation, tests of, 40
Mann, C. R., 115, 121, 129, 131, 145, 150, 151, 162
Mann, Horace, 43, 51
  on written examinations, 29
Man-to-Man scale, Scott, 55
Marconi, Guglielmo, 11
Marking:
  policy, group, importance of, 398–399
  system, need for, 398
  technique:
    pupils' position or rank, 411–412
    scores into marks, 412
Marks, 397–422
  and teachers' popularity, 404
  and teachers' sex, 404
  college freshman, prediction of, 476–479
  definition of marks used, 408–411
  five-letter system, 416
  function of, 399–402
Marks (*Cont.*):
  letter, transmitting point scores into, 413–417
  problem of, 397–398
  proper bases for, 405–408
  "satisfactory" and "unsatisfactory," 412–413
  satisfactory policy, essentials of, 398–411
  sound technique, essentials of, 411–422
  studies in unreliability of, 44–49
  subjectivity of, 45
  transmuting scores into, 412
  weakness in, sources of, 402–405
Marston, William M., 56
Martin, Abe, 19
Martin, Vibella, 481
Massachusetts, written examinations in, 51
Massachusetts teachers Platform for Use of Standard Tests, 177–178
Matching tests, 116, 127, 151–156
  advantages and limitations, 152
  construction, 155–156
  defined, 151–152
  illustrations, 152–155
Mathematical ability, Rogers test of, 41
Mathematics, 6
  grading in, 45
Mathews, C. O., 55
  study of tests, 216
Maxfield, Francis N., 63, 64, 300
Maxwell, James C., 11
May, Mark A., 53, 55
Mayer, Joseph R., 9
Mean, 230–233
  two means, chances that true difference exists between, 250
Measurement:
  errors of, 247–248
    controlling, 12–13
  in guidance, place of, 450–452
  limitations of, 12–13
  means to end, 19
  purposes of, 400–402
  recent trends in, 61–64
  relation of:
    to motivation, 317–319
    to practice in learning, 348
Measurement of Intelligence, The, 36–37
Measuring instrument, satisfactory:
  characteristics of, 65–99
  generalizations, 95–98
  problem, importance of, 65
  reliability of, 65, 82–89
  usability of, 65, 89–95
  validity of, 65–82
Mechanical Ability, General, Stenquist Test of, 41
Mechanical intelligence, 67
Median, 228–230
Medicine, 12
Meece, Leonard E., 258
Mendel, Gregor J., 8
Menninger, Karl A., 311
Mental abilities, primary, 41
Mental age (MA), 226–227, 237, 242
  advantages of, 288
  limitations of, 288–290
  vs. IQ, 287–288
Mental hygiene, 53
Mental imagery rating scale, 55
Mental Measurements Yearbooks, 61, 81–82
Merrill, Maud A., 37, 291, 298, 300, 427

Merrill-Palmer performance scale, 300
Messenger, Helen R., 317, 521
Metropolitan Achievement Tests, 92, 184, 208–209, 303, 311
  errors in arithmetic, analysis of, 372
  sample record, 210
  Supervisors Manual, 310
Meyer, George, 342–343
Meyer, Max, 45
Mid-score, 228
Miles, W. R., 388
Miller, W. S., 289, 297
Mills, Lewis M., 133
Mind, 58
Miner, 56
Minnesota Stabilization and Research Institute, 465
Mitchell, Claude, 334
Mixed dominance, 391
Mode, 227–228
Modern School Achievement Tests, 184, 202, 303
  scoring, 204
Monographs on educational measurement, 60–61
Monroe, Marion, 385, 391
Monroe, Walter S., 49–50, 51, 62, 63, 73–74, 163, 164, 249, 343
Moore, Bruce Victor, 58, 462
Morrisett, L. N., 522, 529
Morrison, Henry C., 290, 348, 435
Morrison, J. Cayce, 200
Morrison mastery formula, 348
Morrison plan, 435
Mort, Paul R., 509–513
Mort-Cornell score sheet, evaluating school system, 512–513
Mosier, Charles I., 145
Motivation, 110, 315–345
  aftereffects of specific connections, 339–341
  and frequency of tests, 322–326
  experiment in, 319–320
    limitation of, 320–321
    types of, 321–322
  implications for:
    educational practice, 344–345
    educational theory, 343–344
  importance of, 315–316
  meaning of, 316
  problem of, 315–317
  relation of measurement to, 316–317
    in learning, 319–343
    in teaching, 317–319
Mountain children, 285–286
Müller, 8
Multiple-choice tests, 127, 145–151
  construction, 150–151
  defined, 145
  illustration, 146–150
  possibilities and limitations, 145–146
Municipal University of Wichita, 478
Murphy, Gardner, 423
Musical Talent, Seashore Test of, 41
Myers, George E., 447
Myers, M. Claire, 145
"Mystery," 403–404

N

National Council of Teachers of English, 80
National Education Association:
  evaluation principles, 495–496
  Sixteenth Yearbook, 488–489
Naval Aviation Cadets, testing, 123
Neale, M. G., 519
Negative learning, danger of, 349–350
Nelson, M. J., 149
Nelson-Denny Reading Test, 188
Nemzek, Claude L., 294
Neumann, 31
Newens, Lyndall Fisher, 160
Newland, T. Ernest, 374
Newman, Sidney H., 378
Newspapers, 517–518
New Stanford Achievement Tests, 276, 277
New-type test, 50
  amateurs and, 104
New York Regents' Examinations, 159, 318
Nineteen Forty Mental Measurements Yearbook, The, 186–188
Noll, Victor H., 325
Nonstandarized tests, achievement, advantages and limitations, 184–185
Nordenskiöld, Erik, 8
Normal curves, 267, 269–270
Normative-survey research, 22
Norms, 282–312
  and standards, 282–284
  comparing intelligence and achievement, 308–310
  educational age (*see* Educational age)
  educational quotient (*see* Educational quotient)
  grade, 304–306
  interpreting scores on:
    achievement tests, 301–308
    intelligence tests, 287–301
    personality tests, 310–312
  percentile, for achievement tests, 306
  personal constant, 298–299
  raw and derived scores, 285–287
  various derived scores, 298–301
  vs. standards, 283–284
North Central Association of Colleges and Secondary Schools, evaluation, 497–498
Norton, John K., 258
Norvell, Lee, 335

O

Oakley, Charles Allen, 465
Objective tests:
  achievement, 50
    origin of, 51
  constructing specific types, 127–156
  frequency of use by teachers, 127–128
  recall types, 127
    completion, 134–138
    simple-recall, 131–134
  recognition types, 127, 139–156
  types of, 127
  validity and reliability of, 128–131
Objectivity and reliability, 88–89
Occupational Titles, Dictionary of, 465
Occupations, 464–472
  shifts in, 466
  United States Census of, 465
Odell, Charles W., 87, 397, 398, 476
Ogburn, William F., 9–10
Ohio State University:
  Psychological Examination, 476, 477, 478
  studies of acceleration, 444
Olson, Helen F., 161
Omar Khayyám, 4

Omwake, K. T., 143
Open-house programs, 529
Opportunities:
  available, analysis of, 453
  educational, analysis of, 473
  vocational, 464–467
Oral tests, 24
Orata, Pedro D., 491
Orator, 19
Organismic philosophy, 17
Organization of school, evaluating, 510–513
Orleans, Jacob S., 441
O'Rourke, L. J., 70, 469
Otis, Arthur S., 40, 244, 269
Otis Scale for rating standard tests, 187, 189
Otto, Henry J., 424, 438, 440, 444
Outside criterion, checking test against, 125

P

Pace, C. Robert, 455, 489, 491
Panlasigui, Isidoro, 329
Parents:
  informal report to, 524
  letters to, 521, 525–527
  reports to, 214
Parent-teacher associations, 529–530
Partridge, E. DeAlton, 440
Paterson, Donald G., 40, 450, 453, 461, 463, 464, 465
Payne, William H., 16
Pearson, Karl, 8, 31, 33, 36, 55, 239
  product-moment coefficient of correlation, 69, 239–241
Pease, Glenn R., 327
Peattie, Donald Culross, 8
Pedagogical Seminary, 59
Pennsylvania State College, profile, 263
Percentile curves, 268–269
  series of distributions, 273–274
Percentile norms, use of, in achievement tests, 306
Percentile scores, 299–300
  standard scores, revised Stanford-Binet IQ's, relation between, 300
Performance, tests of, 40
Personal constant, 298–299
Personal guidance, 482–486
  personality measurement, 482–486
  student problems, 482, 483
Personal interview, 56–58
Personality:
  measurement, 482–486
  of examiner, 96–97
  of teacher, 492
  scores, interpreting, use of norms in, 310–312
  tests, 24
Peters, Charles C., 71–72, 211
Peterson, H. A., 354
Peterson, H. J., 357–358
Peterson, J. C., 357–358
Phillips, Alexander J., 122
Phillips University, 477
Philosopher, scientist and, 17–18
Philosophy:
  of school, evaluating, 498–499
  science and, 14–16
Phrasing of tests, 114–116, 118–119
Physical sciences, 5
  measurement in, 6–7
Physics, 6
Physiology, 8
Pictorial graphs, 254, 258, 270
Pie graphs, 270
Pieper, C. J., 133, 136, 142
Pintner, Rudolph, 40, 50, 200, 290
Pintner-Paterson Performance Scale, 40
Placement, 462
  educational guidance, 481–482
  vocational guidance, 471–472
Planck, Max, 6–7, 96
  on error, 12
Planning tests, teacher-made, 104–112
Platform for Use of Standard Tests, 177–178
Plato, 9, 14, 426
Poffenberger, Albert T., 311
Point scores, 286
Polygons:
  frequency, 265–266, 267
  series of distributions, 272–273
Popular information, 516–517
Positive learning, evidence of, 350–351
Potentialities, 285
Poynter, J. W., 402
Practical subjects, making room for, 51
Practice, 346–363
  final examinations, value of, 361–363
  in correcting examination papers, 355–356
  materials commercially available, 360
  pre-tests, value of, 348–351
  principle of, 346–347
  relation of measurement to, 348
  tests, 50
Praise, 336–339
Pratt, Karl C., 413
Precision instruments, 12
Pressey, Sidney L., 56, 326, 356, 378, 444
Preston, Mary L., 388
Pre-tests, educational value of, 348–351
Preventive diagnosis, 382
  in reading, 393–395
Price, Helen G., 145
Price, Orville Kelly, 69, 70
Primary mental abilities, 41
Principles of Science, The, 30
Prizes, 339
Probability:
  curve, 267
  in guidance, 457–458
  theory of, 10
Problems for which tests are useful, *table*, 181
Proctor, William Martin, 472
Product-moment method, computing coefficient of correlation, 239–241
Professional journals, 58–59
Profiles:
  for series of subjects, 259–262
  individual chart, 260
  of single subject, 259
  use of, in guidance, 262–264
Prognosis tests, 41
Program:
  educational, evaluation of, 499–513
  public relations, 515–533 (*see also* Public relations)
  testing (*see* Testing program)
Progressive Achievement Tests, 18, 184, 284
Progressive Education Association, Eight-Year Study of, 78, 105, 489–490

- Promotion:
  - and retardation, 443–445
  - classification and, 423–446
  - continuous, 445
- Propaganda, 515
- Psycho-galvanometer, 54
- Psychographs, 259
- Psychological examinations, and guidance, 474, 476–478
- Psychological laboratory, first, 30
- Psychology:
  - abnormal, France and, 31–36
  - applied, America and, 36–37
  - correlational, 59
  - experimental, Germany and, 30–31
    - publications, 58–61
  - Gestalt school, 8–9
- *Psychology of Getting Grades, The*, 361
- Psychophysics, 31
- Publications:
  - experimental psychology, 58–61
  - newspapers, 517–518
  - official, 519–521
  - reports, 519–521
  - students, 518
- Publicity, 515
- Public opinion:
  - measuring, 58
  - mobilizing, 530–531
- Public relations, 515–533
  - agencies of public information, 517–518
  - letters to parents, 521, 525–527 (*see also* Letters to parents)
  - mobilizing public opinion, 530–531
  - parent-teacher associations, 529–530
  - programs, 515–517
  - publications, 517–521 (*see also* Publications)
  - report cards, 521–525 (*see also* Report cards)
  - reports to public, 214
  - school exhibits, 527
  - school visitation, 527, 529
- Public School Tests, 184
- Publishers of standard tests, 532–533
- Pullias, Earl V., 104
- Pupils:
  - marking of own papers, 358
  - reports to, 207–208
- Purdue Placement Test in English, 478
- Pure sciences, 5
- Pythagoras, 9

Q

- Quaid, T. D. D., 477
- Qualitative analysis, 8
- Quantitative analysis, 8
- Quartile deviation of scores, 234, 235
- Questionnaire, personality measurement, 54–56
- Quetelet, 10

R

- Range, 233–234
  - of variability, 233
- Rank-difference method, computing coefficient of correlation, 238–239
- Rank order, scores, 219–222
- *Rapport*, and testing, 97
- Raths, Louis E., 106, 107
- Rating scales, 25
  - personality measurement, 54–55
- Ratio, critical, 250
- Raw scores, 285–287
- Raynaldo, D. A., 104
- Reading:
  - clinics, 391
  - comprehension, typewriter graph, 272
  - diagnosis, 382–395
    - case studies, 383–386
    - techniques, 386–395
  - difficulties, analysis of, 388–389
    - causes of, locating, 389–393
  - prerequisites to, 394–395
  - preventive diagnosis in, 393–395
  - readiness, 61
    - scores, classified, 221
    - test, 41
    - class record, 219
  - remedial (*see* Remedial reading)
  - retarded readers, locating, 387–388
  - survey tests in, 387
- Rearrangement tests, 127
- Recall types of tests, 127
  - completion, 134–138
  - simple-recall, 131–134
- Recitation vs. rereading in study, 354
- Recognition types of tests, 127
  - alternative-response, 139–144
  - matching, 151–156
  - multiple-choice, 145–151
- Records:
  - class (*see* Class records)
  - cumulative, *graph*, 213
  - for administrators, 209–214
  - sample, Metropolitan Achievement Tests, 210
  - teachers', 208–209
  - test data summary, N.E.A., 212
  - testing program, 207–214
- Recreational-interests questionnaire, 56
- Reeder, Ward G., 516, 519, 520
- Reeve, Ethel B., 148
- Regression equation, 246
- Relationship, measures of, 237–246
- Reliability, 62
  - coefficient, in evaluating tests, 125
  - differentiated from validity, 83
  - importance of, 83
  - interpretation of, 86–87
  - meaning of, 82–83
  - methods of determining, 83–84
    - with one form, 84–86
    - with two forms, 84
  - objective tests, 128–131
  - objectivity and, 88–89
  - of essay examination, 157–158, 160
  - of measuring instrument, 65, 82–89
  - reputation and, 20–21
- Remedial instruction, tests in, 352–354
- Remedial reading, 386–395
  - procedures, 378–382
  - program, 392
- Remmers, H. H., 76, 326
- Report cards, 398, 521–525
  - Hill's study of, 522–525
  - trends in, 521–522
  - University of Chicago High School, 527, 528
- Reports:
  - annual, 519–520
  - for administrators, 209–214
  - special, 520–521
  - testing program, 207–214
  - to parents and public, 214
  - to pupils, 207–208
  - to teachers, 208–209

Reproof, 336–339
Reputation, 19–20
Rereading, recitation vs., in study, 354
Research, 19, 21–22
Results of tests:
  applying, 205–206
  knowledge of, 329–339
  statistical analysis of, 216–251 (*see also* Statistical analysis of test results)
Retardation, acceleration and, 443–445
Retesting, 206–207
Revised Stanford-Binet Tests of Intelligence, 18
  standard and percentile scores, relation between, 300
Rewards and punishment, 340
Rhetoric, 19
Rice, Dr. J. M., 21, 44, 62
  quoted, 51
Richardson, Helen M., 440
Richardson, M. W., 86
Riley, John L., 52, 520
Rinsland, Henry Daniel, 104, 129, 132, 135, 167
Ritchie, A. D., 4
Rivalry, 338–339
Rock, Robert T., Jr., 340
Rogers, Carl R., 450
Rogers test of mathematical ability, 41
Rosenzweig, Saul, 97
Ross, Clay Campbell, 91, 211, 319, 324, 325, 332, 333, 335, 443, 470
Rothney John W. M., 389
Rousseau, Jean Jacques, 423
R's, three, 19–22
Ruch, Giles M., 23, 60, 80–81, 86, 87, 114, 127, 128, 135, 157, 159, 305, 349, 352, 404, 454, 475, 484
Rugg, Harold, 17, 55
Rundquist, E. A., 478
Russell, Bertrand, 12, 20
  on science in everyday life, 11
  on scientific method, 4–5
Russell, J. T., 243
Russian education, tests and marks in, 28
Ryan, T. M., 147
Ryan, W. Carson, Jr., 158
Ryans, David G., 31

S

Salaries, differences among states, 255
Sales clerks, study of, 465
Sampling, 66
  errors of, 248–249
Sangren, Paul V., 296, 519
Saucier, W. A., 25
Sauvain, Walter Howard, 440
Scale:
  differentiated from test, 25
  first, for measurement of intelligence, 34
  first achievement, 44
Scale-Book, 43–44
Scaled test, 25
Scates, Douglas E., 22, 90, 98, 131, 252, 307, 375
Scattergram, 225–227
Scatter of scores, measures of, 233–237 (*see also* Variability of scores)
Schneidler, Gwendolen G., 450, 453, 461, 464
School administration, function of measurement in, 22–24
School and Society, 59
School plant, evaluating, 510–513
School survey movement, 44
School systems, evaluating, 512–513
Schrammel, H. E., 147
Schutte, 23
Science:
  and philosophy, 14–16
  art and philosophy, 16
  measurement, 3–14
    generalizations, 13–14
  practical applications, 11–12
Scientific method, 4–6
  role of measurement in, 5
Scientist and philosopher, 17–18
Scores:
  knowledge of, and motivation, 329–334
    combined with other incentives, 334–339
  meaning of, 285–286
  on intelligence tests, use of norms, 287–301
  percentile, 299–300
  personal constant, 298–299
  point, transmitting, into letter marks, 413–417
  raw vs. derived, 286–287
  transmuting, into marks, 412
Scoring of tests:
  ease of, 91–92
  errors in, 200–201
  essay examination, 167–171
  procedure:
    rules and answer keys, 123
    teacher-made tests, 122
  reading readiness, classified, 221
  steps in, 204–205
  techniques, 200–202
  who should score, 199–200
Scott, Ira O., 362
Scott Man-to-Man scale, 55
Scruggs, Harry J., 366
Seashore, Carl E., 60, 474, 481
Seashore Test of Musical Talent, 41
Seattle schools, Evaluation Committee, 161
Seay, Maurice F., 258
Secondary school, evaluation, philosophy of, 499 (*see also* High school)
Segel, David, 103, 127, 162, 180, 181, 199, 305, 388, 404, 454, 475, 484
Seguin, Edouard, 40
Selection of tests, 183–193
Self-consistency, 68
Self-criticism, 490
Self-knowledge, 469
Sells, Saul B., 475
Series of distributions, graphs, 272–278
  central tendencies, 274–276
    and variabilities, 276–278
  use of percentile curves, 273–274
  use of polygons, 272–273
Seven Seals of Science, The, 9
Sherman, N. H., 139
Siceloff, L. P., 132, 142
Sigma scores, 300, 306–307
Simple-recall tests, 127, 131–134
  advantages and limitations, 131–132
  construction, 134
  defined, 131
  illustrations, 132–134
Sims, Vernon Martin, 161, 169, 338, 488
Site of school, evaluating, 511, 513
Skewed curves, 269–270
Small-sample theory, 249

Smart, Harold R., 5, 14
Smeltzer, C. H., 358
Smith, B. Othanel, 5, 11, 55
Smith, C. Wilson, 89, 200
Smith, Eugene R., 78, 105, 106, 489
Smith, H. E., 117
Smith, Nila Banton, 352
Smooth curve, 267–269
Social adjustment, 485–486
Social attitudes, Hart's test of, 56
Social intelligence, 67
Social sciences, 5
  genetic history of, 9
  measurement in, 9–10
Sones, A. M., 354
Sones, W. W. D., 132, 155
Soviet Russia, testing and marking in education, 28
Spanney, Emma, 142
Sparling, E. J., 457, 469
Spaulding, Geraldine, 148, 154
Spearman, Charles E., 8, 31, 60, 238–239
  "General Intelligence Objectively Determined and Measured," 58
  rank-difference method, 238–239
Spearman-Brown formula, 84, 85–86, 244–245
Special classes, 442–443
Special talents, 67
Specific intelligence, testing, 41
Specific tests, 61
Speed tests, 113–114
Spelling:
  difficulties and remedies, 379
  Rice inquiry, 21, 44, 51, 62
  variability in, 85
Spence, Ralph B., 175, 399, 413
Spitzer, Hubert F., 354
Springfield, Mass., schools, comparative examinations, 52
Springfield Tests, The, 520
Stalnaker, John M., 158, 159, 168, 169
Standardized tests:
  achievement, advantages and limitations, 184–185
  Otis scale for rating, 187
  publishers of, 532–533
  scales for rating, 187, 189–192
Standards and norms, 282–284
Standard scores (Z-scores), 237, 300, 306–307
  percentile scores, revised Stanford-Binet IQ's, relation between, 300
Stanford Achievement Tests, 184, 303
Stanford-Binet, 36–37, 56, 186
Stanford Revision scale, 36, 67, 69
Stanton, Hazel, 315
Starch, Daniel, 45, 47, 59, 315
Statistical analysis, 58
  methods, England and, 31
Statistical analysis of test results, 58, 216–251
  average or central tendency, 227–233
  classification, 218–227
  error:
    interpretation of, 249–250
    of measurement, 247–248
    of sampling, 248–249
    of technique, 246–247
  importance of statistics, 216–218
  relationship, 237–246
  tabulation, 218–227
  variability or scatter, 233–237
Statistical validity, achievement tests, 70–71
Steeves, H. R., 147
Steiner, M. A., 400
Stenquist, John L., 92, 175, 207, 208
Stenquist Test of General Mechanical Ability, 41
Stephenson, W., 87
Stern, Wilhelm, 31
Stoddard, George D., 60, 133, 194, 297
Stone, C. W., 132, 367
Stone Arithmetic Test, 44
Strang, Ruth M., 388, 392, 451, 462, 475, 480
Strong, Edward K., 469
Strong Vocational Interest Blank, 56
Stroud, J. B., 354
Stuart, Jesse, 360
Student:
  problems, 482, 483
  publications, 518
Studiousness, index of, 308–309
Stullken, Edward H., 374
Subjectivity of school marks, 45
Sullivan, Helen B., 387
Survey-type achievement tests, 50
Sutherland, A. A., 424
Swan, J. N., 284
Symonds, Percival M., 23, 54, 58, 60, 129, 308, 353, 449

T

Taba, Hilda, 317–318
  distribution, 222
Tabulation of test results, 218–227
  frequency table, 222–227
Talents, parable of, 423
Tallmadge, Margaret, 342
Taylor, H. C., 243
Teacher-made tests:
  constructing, 101–171
    objective, specific types, 127–156
  evaluating, 123–125
  frequency of use, *table*, 103
  planning, 104–112
  preparing, 113–120
    author's recommendations, 120
  trying out, 120–123
    normal conditions, 121
    scoring procedure, 122
    time allowance, 121–122
Teachers:
  evaluating, 491–493, 507
  purpose of, and measurement, 317
  records and reports to, 208–209
Teachers College, Examination Inquiry, 158
Teachers College Record, 59
Teaching:
  efficiency, evaluating, 491–493
  evaluating, 509
  relation of measurement to motivation in, 317–319
Teamwork, 339
Temperatures, educational, 500
Tennessee:
  mountain children, 285–286
  teacher-made tests in, 104
Terman, Lewis M., 30, 31, 37, 39, 59, 135, 287, 290, 291, 298, 300, 311, 382, 427, 444
  criteria, 67–68
Terman Group Test of Mental Ability, 69, 201, 297
  scoring, 203

Terry, Paul W., 341–342
Testing program, 172–312
  administering test, 193–199
  applying results, 205–206
  co-operative, 179
  definite, 180–183
  nature of, 175–178
  plan for elementary school, *table*, 176
  practical, 179–180
  purpose of, determining, 178–183
  records, 207–214
  reports, 207–214
  results, statistical analysis of, 216–251 (*see also* Statistical analysis of test results)
  retesting, 206–207
  scoring tests, 199–202
  selecting tests, 183–193
  steps in, 175–215
Test machines, 356–358
Test Newsletter, The, 17
Tests:
  batteries of, 87
  combining, with self-instruction, 357–358
  companies issuing, 189
  construction of, 61–62 (*see also* Construction: tests)
  cost of, 93–94
  differentiated from scales, 25
  ease of scoring, 91–92
  frequency of, and motivation, 322–326
  interpretation of, 92–93
  mechanical make-up of, 94
  on college level, 23
  scaled, 25
  truthfulness of, 65–66
  types of, 24–26
  use of, 62–64
Thermometers, 501
Thinking, critical, 108–109
Thompson, Clarence C., 337
Thorndike, Edward L., 17, 30, 36, 37, 45, 59, 60, 96, 207, 267, 316, 337, 339, 340, 341, 346, 347, 479, 480, 490, 491
  father of educational measurement, 44
  1918 paper, 49
  on tests, 73
Thorndike Handwriting Scale, 44
Thought questions, 163–164
Three R's, 19–22
Three-Year Study of Committee on Teacher Education, 489
Thurstone, Louis L., 41, 104, 290, 301, 469
  attitude questionnaires, 56
Thurstone, Thelma Guinn, 301
Tiegs, Ernest W., 398
Time allowance for tests, 121–122
Time for administering tests, 193–195
Time-sampling, 58
Time Series Charts: A Manual of Design and Construction, 280
Todhunter, Isaac, 20
Toops, Herbert A., 78, 464
Trait variability, 428–429
Traxler, Arthur E., 148, 160, 176, 189, 193, 211, 214, 360, 370, 388, 392, 454, 482, 493
  handwriting, difficulties and remedies, 380–381
  high school reading tests, comments on, 188–189
Trial and error, 4
Trimble, Otis C., 76
Trowbridge, Margery Hayden, 341
Troyer, Maurice E., 455, 489
True-false tests, 139–144
  phrasing of, 114–115
Truthfulness of test, 65–66
Trying out tests, teacher-made, 120–123
T-scores, 237, 306–307
Tucker, A. C., 209
Turney, Austin H., 23, 315, 324, 438, 440
Two-way frequency table, 225–227
Tyler, F. T., 328
Tyler, Ralph W., 78, 105, 106, 374, 489, 516
  on achievement-testing movement, 21–22
  on validity of achievement tests, 74–76
Typewriter, bar graph made on, 270, 271

U

Ungrouped series, scores, 219
United States Employment Bureau, 465
United States Office of Education, 480
Universities, evaluation, 497–498
University of Chicago High School, 160
  marks of students, 45
University of Kentucky, 335
University of Minnesota, 325
  achievement tests, 105
  guidance program, 449–450
University of Missouri, study of marks, 45
University of West Virginia, 158
Updegraff, Ruth, 194
Upton, Clifford B., 80
Urges, 316
Usability of measuring instrument, 65, 89–95
  cost, 93–94
  ease of administration, 90–91
  ease of interpretation and application, 92–93
  ease of scoring, 91–92
  essay examination, 160
  meaning of, 89–90

V

Validity:
  achievement tests, 70–82
    construction of, 75–79
    criticisms of validity, 73–74
    curricular vs. statistical validity, 70–71
    direct vs. indirect methods, 76–78
    item validity, 78–79
    methods, 71–73
    standard tests, judging, 80–82
    Tyler's suggestions, 74–76
  coefficient, 69, 244
  differentiated from reliability, 83
  essay examination, 157, 160–161
  evaluating tests, 124
  general considerations, 66
  intelligence tests, 66–70
    meaning of intelligence, 67
    Terman criteria, 67–68
    individual vs. group tests, 68–70
  meaning of, 65–66
  measuring instrument, 62, 65–82
  objective tests, 128–131
Van Omer, Edward B., 264

Variability:
  human (*see* Human variability)
  meaning of, 233
  of scores:
    quartile deviation, 234, 235
    range, 233–234
    standard deviation, 234–237
    which measure is best, 237
  trait, 428–429
Variations:
  in central tendency, marks, 417–421
  index of, evaluation, 513–514
  in IQ, 6 tests, first-grade pupils, 296
Verification, 18
Versatility, 428–429
Veterans:
  schools for, 474, 476
  tests for, 107
Vineland, N. J., 36
Visitation, school, 527, 529
Vocational guidance, 53, 463–472
  counseling, 470
  follow-up, 471–472
  inadequacy of program, 463–464
  individual, analysis of, 467–469
  opportunities and requirements, analysis of, 464–467
  placement, 471–472
  technique, 453
Voelker, 53
Von Borgersrode, Fred, 189
Votaw, David F., 119

W

Walker, Helen M., 493
Washburne, John Noble, 258
Washington, George, 20
Watson, Goodwin, 54, 93, 339
  measurement of fairmindedness, 56
Watts, Winifred, 521
Weber, E. H., 8, 31
Webster, Edward C., 463
Weidemann, Charles C., 159, 160, 162, 163, 168
Weitzman, Ellis, 123, 146, 150
Wells, F. L., 290
Wesley, E. B., 148
Westaway, F. W., 6, 12, 13, 95
Wheeler, L. R., 285, 286
White, Clyde W., 318
White, Emerson E., 29
White, Hubert B., 328–329
Whitehead, Alfred N., 4, 8, 15
Wilbur, R. L., 375
Wilder, M., 325
Wilds, Elmer Harrison, 442
Williams, Clarence O., 264
Williams, J. Harold, 281
Williams, L. A., 401, 439
Williamson, E. G., 402, 450, 453, 461, 463, 464, 470, 482
Wilson, Guy M., 98
Winnetka plan, 430–432
Wisconsin University, 478
Wise, Virgil, 413
Wissler, Clark, 36
Witty, Paul, 56, 388, 391, 393, 443
Wood, Ben D., 60, 93, 147, 211, 284, 425 481
Woods, Gerald G., 355
Woodworth Personal Data Sheet, 55
Woody, Clifford, 97, 519
Woody, Thomas, 435
Worcester, D. A., 166
Word blindness, 391
World success, 67, 68
World War I:
  Army tests, 40–41
  Scott Man-to-Man scale, 55
  Woodworth Personal Data Sheet, 55
World War II, 41
  accelerated programs during, 444
Wrenn, C. Gilbert, 189
Wright, W. H. E., 139
Wrightstone, J. Wayne, 77–78, 149, 153, 165, 392
Wrinkle, William L., 526
Wriston, Henry M., 490
Written examinations, 24, 51 (*see also* Essay examinations):
  value of, 29
Wundt, Wilhelm, 30–31, 32, 36
Wyndham, Harold S., 439

X

X-O Test, 56

Y

Young, Paul Thomas, 316

Z

Zeigfeld, Edwin, 491
Zook, G. F., 497
Z-scores, 237, 300, 306–307
Zubin, Joseph, 339

Date Due

Wyckoff			
Curriculum			
~~Reserve~~			
MY 14 '54			
~~FACULTY~~			
~~JAN 8 [illegible]~~			
~~FEB 6 [illegible]~~			

LB1131 .R82
Measurement in today's schools
Princeton Theological Seminary–Speer Library
1 1012 00064 2910